Citation for published version:
Tudhope, D, Koch, T & Heery, R 2006, *Terminology services and technology: JISC state of the art review*. Joint Information Systems Committee.

Publication date:
2006

Link to publication

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 06. Nov. 2019

Terminology Services and Technology

JISC state of the art review

Douglas Tudhope — University of Glamorgan
Traugott Koch — UKOLN, University of Bath
Rachel Heery — UKOLN, University of Bath

Document details

Date:	15-09-2006
Version:	Final draft for approval
Notes:	Circulation to JISC Development Team

Acknowledgement to funders

This work was funded as part of the JISC Information Environment.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

Contents page

Executive Summary

Purpose

Over the next two years, as part of its Capital Funding Programme, the Joint Information Systems Committee (JISC) is supporting further work to realize a rich information environment within the learning and research communities. This review is intended to inform JISC's planning for future work related to Terminology Services and Technology, as well as to provide useful background information for participants in future calls, whether specifically featuring terminology or where terminology can be used to underpin other services.

Overview of report contents

This report reviews vocabularies of different types, best practice guidelines, research on terminology services and related projects. It discusses possibilities for terminology services within the JISC Information Environment and eFramework.

Terminology Services (TS) are a set of services that present and apply vocabularies, both controlled and uncontrolled, including their member terms, concepts and relationships. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc. Indicative use cases are discussed.

One type of TS attempts to increase consistency and improve access to digital collections and Web navigation systems via vocabulary control. Vocabulary control aims to reduce the ambiguity of natural language when describing and retrieving items for purposes of information searching. Another type of TS is not concerned with consistency but with making it easier for end-users to describe information items and to have access to other users' descriptions. This results in vocabularies (folksonomies) that may not be controlled, at least initially. The report reviews different kinds of vocabularies, according to their structure and their intended purpose. Potential benefits and return on investment are discussed. Named entity authority and social tagging services are discussed in some detail. Pointers are given on best practice guidelines and networked access to vocabularies, including key issues for future terminology registries.

The wider context of TS is considered. Relevant literature on user studies is reviewed. TS are located within an information lifecycle and within the JISC IE. Suggestions are made towards a more specific definition of Terminology Web Services within the JISC IE. Current work on Terminology Web Services is reviewed, along with work on mapping, automatic classification/indexing and repositories. Current projects that involve TS activity (JISC, UK, and international) are briefly reviewed.

Relevant standards are discussed, particularly for vocabulary representation; identification of concepts, terms and vocabularies; protocols and APIs.

Key points

TS can be m2m or interactive, user-facing services and can be applied at all stages of the search process. Services include resolving search terms to controlled vocabulary, disambiguation services, offering browsing access, offering mapping between vocabularies, query expansion, query reformulation, combined search and browsing. These can be applied as immediate elements of the end-user interface or can underpin services behind the scenes, according to context. The appropriate balance between interactive and automatic service components requires careful attention.

Return on investment should be considered in any service provision. There are various types of vocabularies serving different purposes, with different degrees of vocabulary control, richness of semantic relationships, formality, editorial control. There are a range of TS options, both interactive and automatic. There is potential for piloting TS to augment existing JISC programmes and projects.

TS are sometimes contrasted with free text searching, assisted by statistical Information Retrieval techniques in automatic indexing and ranking. These are not, however, exclusive options and there are opportunities in exploring different combinations of the two approaches. It should be noted that Web search engines have introduced elements of TS, by offering synonym and lexical expansion options. Thus TS should not be seen as antithetical to free text searching and can augment it.

There are many existing vocabularies. Different arrangements regarding ownership, maintenance and licensing of vocabularies can be found. The issue of who will maintain a vocabulary and the basis on which it can be described or made available in a registry needs investigation since this underpins systematic use of vocabularies in the JISC IE. This involves establishing business models for access to and maintenance of vocabularies.

Mapping is a key requirement for semantic interoperability in heterogeneous environments. Although schemas, frameworks and tools can help, detailed mapping work at the concept level is necessary, requiring a combination of intellectual work and automated assistance. The impact on retrieval is a key consideration.

Automatic classification and indexing tools are important for addressing the potential resource overheads in applying TS to indexed collections and repositories. Some tools are emerging that should be investigated for JISC purposes. Many argue for a combination of intellectual and automatic methods.

It is important to consider how people search for information when designing and evaluating TS, in order to reduce the scope for design errors and increase the possibility that services will actually be used. User studies should be conducted where feasible in ongoing project work.

TS should not be seen as an isolated, free-standing component. TS need to be considered within the wider context of the JISC IE, and need to be integrated with other components

of the eFramework. They should be seen as forming a set of services that can be combined with a wide range of other services. There is a need for specifications of TS and their workflow, as part of the JISC IE.

Interoperability requires commonly agreed standards and protocols. Standards exist at different levels and types of interoperability. The prospect is emerging for a broad set of standards across different aspects of terminology services - persistent identifiers, representation of vocabularies, protocols for programmatic access, vocabulary-level metadata in repositories. Such standards are an infrastructure upon which future TS will rest but it is not feasible to wait for international agreements; international consensus will be influenced by operational experience. Pilot TS projects should orient to existing potential standards (in persistent identifiers, representations, protocols for programmatic access) and help to evaluate and evolve them.

Recommendations

The review was asked to include: “recommendations for further activities needed in this field, and the extent to which JISC should be involved in the work (both short and longer term), including collaboration with other organizations as a possible form of involvement". The following recommendations are listed according to the relevant section of the review, where further context may be found.

1. Introduction

1.1 Purpose of this review

- Terminology services can support various stages of the information lifecycle
- JISC should highlight subject access and terminology services in all relevant JISC programmes, whether as extensions to existing projects or as new projects

1.2 Terminology Services overview

- Demonstrate integration of Terminology Services with other components of the JISC Information Environment. (See also Recommendation 4.3)

1.2.3 Combination of terminology tools and techniques

- Encourage inter-disciplinary collaboration in the development of terminology services and co-operation with memory institutions and archives
- Investigate different combinations of TS and uncontrolled (non-TS) search

1.3.2 Return on investment

- Investigate methods to make vocabularies available to the education sector through a Registry, initially for experimentation purposes but ultimately in a sustainable, maintained, licensed manner. (See also Recommendation 3.7)

2 Use cases - scenarios

- Use cases should be developed and refined in an ongoing basis, along with case studies of TS in practice, user session logging, observation, etc.

3 Types of vocabularies

- Provide access to a range of different vocabularies according to context
- It is important to consider the broader context and return on investment

3.1 Vocabularies by structure

- Consider faceted approaches when developing vocabularies and TS

3.2 Vocabularies by purpose

- Descriptions of intended purposes of a vocabulary would be a useful element of a vocabulary registry (see also Recommendation 3.7).

3.2.4 eLearning purposes

- Increased cross-fertilisation between eLearning and Digital Library fields
- User studies of behaviour by indexers (cataloguers), students, teachers. Investigate how to support effective practice with a variety of indexing and retrieval tools
- Investigate conversion between VDEX and SKOS Core representations for compatible vocabularies (see also Recommendation 6.2).

3.2.5 eScience purposes

- Studies of user practice with vocabularies describing research data

3.3 Named entity authority and disambiguation services

- Investigate lists of institutional names and academic affiliations (IESR Agents etc.)
- Study the coverage of available name authorities in OPACs and academic web publishing (LEAF, CiteSeer and similar)
- Engage in international cooperation (eg, LEAF, OCLC, SURF DARE)
- Prototype a demonstrator UK Name Authority File, possibly involving BL and universities (authentication, staff, institution databases) and evaluate its use in a limited application
- Address the treatment of place and geographical names in UK services and activities, and the development of standards and authorities, in cooperation with related projects and terminology efforts.
- Support active participation of UK institutions in international naming standardisation efforts in scientific disciplines and, via project support, assist their implementation in UK
- Apply methods of name extraction and investigate their benefits compared to and in combination with traditional authority systems. Build and evaluate different name disambiguation demonstrators
- Experiment with a Name Authority Web Service, e.g. to be built into metadata creation tools
- Develop or support metadata enhancement services for correction and enrichment: vocabularies, schemes, mapping, names

3.4 Social tagging and folksonomies

- Experiment with combination of KOS-based controlled indexing with an established vocabulary and free (social) tagging for research purposes in a specific discipline, optimised for discovery and retrieval
- Experiment with potential for automatic linking of tags to facets, controlled vocabularies and authorities
- Integrate tagging with existing services such as repositories, OPACs, (RDN/Intute) subject gateways, Digital Libraries, KOS creation and management systems, museum exhibitions and catalogues, metadata enhancement services etc.
- Comparison study between different types of user participation: annotation, recommendation, personalization, restructuring of information, categorization, concept space, concept maps, topic map tools. This could inform a prototype integrating different types of user participation with social tagging

3.7 Terminology Registries

Demonstrate the use of a terminologies registry within JISC IE testbed to include

- Investigating inclusion of terminologies into IESR, potentially describing vocabularies as collections
- Developing marketing proposition for a UK terminology registry (include use scenarios, IPR issues, business models, cost benefit)
- Evaluating use of the draft metadata description profile proposed by NKOS
- Maintain collaboration between various UK initiatives (with eScience e.g. GRIMOIRES and learning communities e.g. Becta Vocabulary Tool) and internationally (e.g. NSDL)

4 Activities with TS

4.1 Studies and models of information seeking behaviour

- User studies of TS in context of JISC IE, illuminating the search process (for work flow of services) and the appropriate balance between interactive and automatic TS

4.3 Types of Terminology Web Services

- Develop more precise definitions of TS, as part of the JISC IE and eFramework
- Define search process workflow of TS within JISC IE eFramework
- Within the context of eFramework, develop a hierarchical layered set of protocols for TS and standard bindings to (various) APIs
- Develop open source, reference terminology web service implementations

4.3.4 Terminology Web Services review

- Collaborate with international efforts in terminology web services
- Develop a range of TS-based search and browsing tools

4.4 Mapping

- Investigate/compare different mapping approaches and granularities in pilot projects
- Develop a range of TS-based tools to assist in creating mappings

- Investigate the potential for standard mapping relationships and a mapping protocol
- Collaborate with international efforts in mapping services

4.5 Automatic classification and indexing

- Investigate semi-automatic solutions to indexing and classification in pilot projects
- Investigate currently available tools for automatic indexing and classification

4.6 Text mining and information extraction

- Investigate relationship between KOS and text mining:
    - Demonstrate how KOS can support text mining
    - Demonstrate how text mining can be used to update and enhance KOS

5 Review of current terminology service activity

- JISC should negotiate Dewey licenses for JISC services and projects

5.5 Repositories

- Pilot different approaches to subject based access to repository content via different types of vocabulary and TS, taking cost benefit issues into account and various levels of aggregation of content:
    - use of subject classification and
    - use of specialised KOS vocabularies
    - use of author assigned keywords
    - full text indexing
- Consider use of mainstream classification (such as DDC) in combination with assigning specialised vocabulary terms (as in use within RDN)

5.6 Augmenting existing programmes and projects

- JISC should support a range of pilot demonstrators with end-users and evaluation
    - Investigate different TS approaches to (eg) indexing, mapping, search/browsing, query expansion, disambiguation
    - Consider subject access and terminology service adjuncts to appropriate JISC programmes and projects, including TS support for Intute; connection of TS (and subject access) to collection level metadata (e.g. topical composition, correlation); TS support for repositories; project-specific examples
- Harvesting
    - Investigate possibilities for extending harvesting tools with more subject metadata
    - Investigate relationship of TS and OAI etc
    - Evaluate benefits of vocabulary-oriented metadata normalising and enhancement service, e.g. aggregator harvesting relevant metadata, enhancing it and then offering harvesting of the improved metadata
- Develop vocabulary visualisation tools supported by TS
    - Flexible display and tailoring of segments from vocabularies

- Flexible display and tailoring of results
- Combined search/browsing

6 Standards

- JISC should encourage participation in international standardisation activities

6.1 Design

- Relevant standards should be included in JISC Standards Catalogue. All new initiatives should take account of relevant design standards

6.2 Representations

- Strongly recommended to use XML-based representations
- Recommended that vocabulary providers consider using SKOS Core if appropriate and contribute to further extensions and customising of SKOS Core

6.3 Identification of concepts, terms and vocabularies

- A global identifier mechanism for referring to vocabularies and their components underpins interoperable TS
- Recommended to consider building upon existing work with the http URI approach for concept identifiers
- Investigate the addition of identifiers to a widely used freely available vocabulary in a pilot study
- Educational work with vocabulary providers on need to supply identifiers and discussions on practical issues should be undertaken

6.4 Protocols, profiles and APIs

- Need for standard m2m protocols for networked access to vocabularies (and their constituent concepts, relations and terms) with common bindings (APIs) building on web services and other low-level standards
- Recommended to consider using SKOS or ZThes API for TS (with a view to contributing to further development). Investigate possibilities of unifying SKOS and ZThes APIs
- Investigate possible standard m2m protocols for mapping access to vocabularies, perhaps by expanding SKOS or ZThes APIs
- Investigate the combination/integration of TS with existing query APIs (SRU/SRW, CQL) or possibly develop new TS-based query APIs

1. Introduction

1.1 Purpose of this review

Over the next two years, as part of its Capital Funding Programme, the Joint Information Systems Committee (JISC) is supporting further work to realize a rich information environment within the learning and research communities. This review is intended to inform JISC's planning for future work related to Terminology Services and Technology, as well as to provide useful background information for participants in future calls, whether specifically featuring terminology or where terminology can be used to underpin other services. The review is intended to identify useful areas of activity and highlight current initiatives of interest rather than be comprehensive or prescriptive. The review will recommend a number of areas with potential, either for further investigation, or for the development of tools or demonstrator services.

JISC's interest in terminology services is part of its strategy for shared infrastructure services underpinning resource discovery, both m2m and user-facing services. Within the education sector, there is interest from service-provider, developer and service user representatives. With adoption of a Services Oriented Approach (SOA), there is potential for a granular approach to terminology services, with different services arising from and being maintained by different communities. There is potential for re-use of some widely-applicable services to support learning, teaching and research, with other providers (possibly including other public sector bodies, research communities, professional societies and so on) providing more specifically-focused services.

The recommendations made by this report are based both on a review of current activity and on contacts made with a number of interested parties. There is some overlap of topic with the JISC Shared Infrastructure Services Review and the JISC Pedagogical Vocabularies Project. This review has a more specific focus on Terminology Services, making reference to the other studies as appropriate. Multilingual vocabulary support, translation support, spelling correction and dictionary services are considered out of scope. References are provided at the end of the report, organized by section.

Recommendations:
Terminology services can support various stages of the information lifecycle.
JISC should highlight subject access and terminology services in all relevant JISC programmes whether as extensions to existing projects or as new projects

1.2 Terminology services overview

Terminology Services (TS) are a set of services that present and apply **vocabularies**, both **controlled** and **uncontrolled**, including their member terms, concepts and relationships. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc. They can be m2m or interactive, user-facing services and can be applied at all stages of the retrieval process.

TS can be confusing in that they span very different application areas, vocabularies, communities, and can provide quite different kinds of services. They can be applied as immediate elements of the end-user interface (e.g. pick lists, browsers or navigation menus, search options) or can underpin services behind the scenes.

TS need to be considered within the wider context of the JISC Information Environment, and need to be integrated with other components of the environment and with other services (Section 4.3). Standard representations, protocols and APIs need to be defined to enable programmatic access and encourage interoperability (Section 6).

Recommendation: Demonstrate integration of Terminology Services with other components of the JISC Information Environment. (See also Recommendation 4.3)

Vocabularies are often associated with control of subject (or topic) metadata. This includes the major bibliographic or educational subject classifications, thesauri used for subject indexing, species taxonomies, etc. Other types of metadata can also benefit from vocabulary control, prominent examples including place names, personal names, genre and various descriptors of educational context in eLearning.

1.2.1 Controlled vocabularies

One type of terminology service attempts to increase consistency and improve access to digital collections and Web navigation systems via vocabulary control. Vocabulary control aims to reduce the ambiguity of natural language when describing and retrieving items for purposes of information searching.

Controlled vocabularies consist of **terms**, words from natural language selected as useful for retrieval purposes by the vocabulary designers. A term can be one or more words. A term is used to represent a **concept**.
Two features (synonyms and ambiguity) in natural language pose potential problems.

a) Different terms (synonyms) can represent the same concept.
b) The same term (homographs) can represent different concepts.

A **controlled vocabulary** can attempt to reduce ambiguity between terms by :-
- defining the scope of terms - how they are to be used within a particular vocabulary.
- providing a set of synonyms or effective synonyms for each concept
- restricting scope so that terms only have one meaning (and relate to only one concept).

Not all vocabularies provide all three features above. Some are just simple lists of authorized terms (authority lists). Controlled vocabularies also provide vocabulary for **Knowledge Organization Systems (KOS)**, which additionally structure their concepts via different types of semantic relationship. Types of KOS are discussed in Section 3.

1.2.2 Folksonomies

Another type of TS is not concerned with consistency but with making it easier for end-users to describe information items and to have access to other users' descriptions. This

results in vocabularies that may not be controlled, at least initially. In principle, this is not a new type of terminology but novel web applications have gained attention recently. Various neologisms have emerged for this activity, including **social tagging** and **folksonomies**. It is seen by some to hold promise of reducing indexing costs and perhaps most significantly, encouraging end-user participation in information services and contributing to community building. However it has yet to be evaluated for educational purposes and existing social tagging applications have not been designed with general retrieval in mind. Folksonomy-based terminology services are discussed in Section 3.4.

1.2.3 Combination of terminology tools and techniques

TS are sometimes contrasted with free text searching, assisted by statistical Information Retrieval techniques in automatic indexing and ranking. These are not, however, exclusive options and there are opportunities in exploring different combinations of the two approaches. It should be noted that Web search engines, such as Google, have introduced elements of TS, by offering synonym and lexical expansion options. Thus TS should not be seen as antithetical to free text searching and can augment it.

In general, different disciplines make use of vocabularies and can contribute to TS, including Artificial Intelligence, Human-Computer Interaction, Information Retrieval, Library & Information Science and Natural Language Processing.

Recommendations:
Encourage inter-disciplinary collaboration in the development of terminology services and co-operation with memory institutions and archives
Investigate different combinations of TS and uncontrolled (non-TS) search

1.3 Cost benefit issues

Various cost benefit issues relating to terminology services should be considered.

1.3.1 Benefits

Terminology Services enable users to undertake educational and research inquiries more effectively. When searching free text with uncontrolled terms, significant differences can stem from trivial variations in search statements and from differing conceptualisations of an information need. Different people use different words for the same concept or employ slightly different concepts. It can be difficult for non-specialists to employ technical vocabulary and variation in person or place names can frustrate consistent access. This may not be a problem if the purpose is just to obtain a few relevant items as examples of a topic. However, when the purpose is an in-depth educational review or systematic research on a specialized topic then it is undesirable to miss potentially relevant items. These problems can be helped by various Terminology Services.

At the simplest level, a controlled list of terms ensures consistency in **searching and indexing**, helping to reduce problems arising from synonym and homograph mismatches. **Name authorities** are an important example.

At a more complex level, the presentation of concepts in hierarchies and other semantic structures helps the indexer and searcher choose the most appropriate concept for their purposes. **Browsing** based user interfaces become possible.

A KOS can assist both precision (by allowing specific searching) and recall (by retrieving items described by related concepts or equivalent terms). It also provides **potential pathways (for human and machine)** that connect a searcher and indexer's choice of terminology. The more formal specification of specific semantic relationships in an ontology can assist applications where rules are specified on the relationships and logic-based inferencing is appropriate.

The use of uncontrolled vocabularies may encourage end-user participation in **social indexing or tagging** and help build user communities for an application.

Many **mapping and semantic interoperability** applications depend upon KOS of different types, as do other downstream applications.

1.3.2 Return on investment

The return on investment (ROI) should be considered. There are many different kinds of vocabularies, with different degrees of vocabulary control, richness of semantic relationships, formality, editorial control - all serving slightly different purposes (see Section 3).

Different arrangements regarding ownership, maintenance and licensing of vocabularies can be found. The issue of who will maintain a vocabulary and the basis on which it can be made described or made available in a registry needs investigation since this underpins systematic use of vocabularies in the JISC Information Environment. This would involve establishing business models for access to and maintenance of vocabularies.

Recommendation:
Investigate methods to make vocabularies available to the education sector through a Registry, initially for experimentation purposes but ultimately in a sustainable, maintained, licensed manner. (See also Recommendation 3.7)

There is overhead in designing a controlled vocabulary and also in its use for classification or indexing. Thus cost/benefit issues should be considered for the particular application in mind, when deciding on richness of semantic relationships and degree of formality. For example, is a simple controlled authority list sufficient for the purpose? On the other hand, there are many existing vocabularies and indexed datasets which can be leveraged or combined in larger schemes. There is also potential in (semi)automatic indexing and classification techniques, both in application of products from commercial systems and outcomes of projects in this area (see Section 5). There is also potential in the application of interactive metadata assignment tools and their embedding in application interfaces and project workflow.

2 Use cases - scenarios

In light of the possibilities discussed in this review, there is general potential for TS augmentation - some near term, some longer term. The following high-level scenarios and discussion are intended to illustrate a selection of the benefits TS might offer.

Some of the following scenarios are based on the RDN (recently re-launched as Intute) as an example of a prominent information service that could be further improved by various TS, there is no intended criticism as similar points could apply across many JISC information services

2.1 Retrieval performance

Your teacher has given an assignment to find information from the RDN on how *vog* is relevant to tomorrow's classes. Unfortunately your attention wandered momentarily at the point when this new word was explained. You do not know if it is something to do with the morning class on Japanese culture and street style or the afternoon's class on volcanos and global warming. You do a search with RDN on *vog* and find no hits. Using a TS that searches a general subject vocabulary, you look up *vog* and find it is related to *volcanic gases*. You search RDN with these terms and find relevant resources

This scenario illustrates a hypothetical web service that suggests extra terms to construct or refine a query. This scenario is an abridged version of Vizine-Goetz's scenario for OCLC Research Terminology Services which involved a general Web search http://www.oclc.org/news/publications/newsletters/oclc/2004/266/research.html. That scenario employed Library of Congress Subject Headings, as an example of an authoritative and frequently updated general vocabulary. LCSH is sometimes used as a general classification scheme, often along with more specialized vocabularies. Similar TS could be offered by other vocabularies, in more specialized domain applications.

Initial stages of a search process (see Section 4.1) may involve a process of exploration or familiarization with details of an information need. As well as general subject vocabularies, online dictionaries or encyclopedia are sometimes used for this general purpose. Various TS could be integrated as an option in the search process, as sources for query terms. Google Toolbar already offers a dictionary service and similar forms of TS can be envisaged.

In information retrieval systems, Synonym Rings or Search Thesauri (see Section 3) are used for purposes of improving search performance by taking account of synonyms and also terms from related concepts in matching a query. A range of TS services to improve query performance (both recall and precision) are possible. This includes various query expansion possibilities, where result ranking can be based on degree of semantic match. For example, you may wish to search with very specific terminology; you would be very interested in matches on those concepts and, failing that, would also be interested in matches on closely related concepts. Employing query expansion can combine several search 'moves' in the one query.

Another example scenario from OCLC Research is an item from Dempsey's Weblog (Aug 18, 2005) on how the catalog can be used to offer access paths, via *same author* or *similar topic*, etc. This is demonstrated with an example from the Worldcat *Find in a Library* service. http://orweblog.oclc.org/archives/000772.html

2.2 Name Authorities

You wish to find articles by author *D. Smith* in your ePrints University repository. This allows search by *Smith, D*. However, a large number of search results are returned, with several variants of the name (including hyphenated surnames, first names, middle initials), representing several, different authors, in a single list. There is no easy way to disambiguate the different people and achieve a definitive list. There is no online authority file which you could search or browse and select the definite person. The situation becomes even more difficult when the author name occurs as both first name and surname (eg *Thomas*, or *Michael*).

Provision of an integrated 'added value' name authority service would allow the searcher to disambiguate author's names.

2.3 Mapping and other TS

This scenario extends the attractive BIOME Alternative Land Uses Case Study by considering mapping and some other TS:

> Farm diversification is often held up as a panacea for a time of falling prices at the farm gate. Changing to new farm products or going into organic or conservation grade production is viewed as on route out of the cycle of downward farm gate prices. But also environmental schemes often referred to as agric environment schemes are put forward as a way of stabilising farm incomes and giving benefits to the wider community both rural and urban. http://www.rdn.ac.uk/casestudies/biome/agriculture/case4.html

The scenario discusses two AgriFor resources, resulting from a BIOME search on *farm environment schemes.* The query is given as a starting point in the case study. However it is not obvious that a student would formulate such a query as a first step. Various TS might help in constructing the query by suggesting controlled terms (as discussed in Section 2.1), or with different forms of query expansion – both synonym and concept expansion.

Initial browsing is also a common early stage of the search process. Browsing is available by AgriFor high level categories and relevant items can be found under *Economics, Trade and Rural Development/ Agricultural Economics/ Government farm policies.* The third level category is not visible on the main Browse screen, so a student would need to select *Agricultural Economics* when browsing. See Section 2.4 for a discussion on a vocabulary search TS for extended browsing systems.

Having found the information item mentioned in the case study, there is no easy way of 'beaming-up' to the AgriFor categories, other than the browser back button. Instead, information items are indexed with CAB Thesaurus concepts. This is helpful – knowing

how items are indexed is potentially useful for refining a search. The CAB concepts provide an option for navigating through the collection via clicking on an index descriptor. However, the thesaurus structuring of knowledge is not available to further assist the search. Browsing is via an alphabetical list - hierarchical context and related concepts are not available.

Combining a classification with a thesaurus for indexing provides excellent resources. More use could be made of the combination. One possibility is to map the two vocabularies together. This might help advanced search facilities, such as query expansion. Another possibility is a greater integration of search and browsing (see the DeweyBrowser, Section 5.3.6).

At present, the RDN case studies tend to be isolated within one of the BIOME gateways. Mapping could extend beyond the two vocabularies used inside AgriFor to the BIOME vocabularies, generally. For example, the Natural Selection gateway also contains useful resources for the case study. Natural can be browsed by DDC headings and information items are indexed by free-standing keywords. The collection includes items on Computer Based Learning in Land Use and Environmental Sciences, a journal on natural resource management and restoration, technology for ecology management, DEFRA wildlife and countryside, etc. – all potentially relevant to the case study. A mapping between the DDC headings, AgriFor categories, CAB Thesaurus could underpin a variety of TS and access routes. Cross browsing and cross-searching would be enabled across the two collections.

2.4 Repositories

You wish to search your institutional ePrints repository for articles on a particular subject. Since the coverage is wide, a general vocabulary is available for browsing access, in this case the top 2-3 levels of the Library of Congress Subject Areas, with associated postings. However, it is not clear from the main menu where your subject interest would fall – the terms you usually employ to describe your subject are not mentioned and you don't feel like browsing multiple sub-menus in the quite extensive browsing classification. In the browser, you try to *Find on this page* without success. There is no way of searching the vocabulary to find where your interest might fall. You can, of course, search the full text but this relies on a subject keyword appearing in the text.

A TS that augmented the general classification with an *entry vocabulary* of synonyms and allowed search of this extended vocabulary would extend the utility of the retrieval functionality. This would provide additional entry points for browsing. The more extensive the classification and the browsing options, the more useful this will be.

This scenario assumes that subject search of a University publication repository is a sensible option. Given the probable patchy distribution of coverage in any one University, some form of known item search or author-based search may be more likely. However, subject-based access would be applicable to various types of aggregated repositories in the future.

Recommendation: Use cases should be developed and refined in an ongoing basis, along with case studies of TS in practice, user session logging, observation, etc.

3 Types of vocabularies

Descriptions and comparisons of different types of vocabularies are often confusing because the terminology is not controlled and there is also a fair degree of overlap. Furthermore, systems can be compared across different criteria. For example vocabularies differ in structure and levels of complexity but also in the application purposes for which they are designed and used. We first consider vocabularies by their structure and then discuss them according to some major high level purposes or application areas.

Recommendation: Provide access to a range of different vocabularies according to context

3.1 Vocabularies by structure

Vocabularies can be considered by their structure (Hodge 2000 and see also BSI, NISO). One way of organizing them is by increasing structural complexity and types of relationship, which is roughly the order of the main divisions in the following discussion. Knowledge organization systems (KOS) are controlled vocabularies, which are organized and structured via different types of semantic relationships.

3.1.1 Term Lists

At the simplest level, **Term Lists** offer *ambiguity control* and, usually unstructured lists, are particularly appropriate when a limited set of options is offered. If made available as a **pick list**, they can ensure terminology control in interactive indexing and searching applications. **Authority Files** are used to control variants of named items, such as personal, organizational or place names, and are often presented in alphabetical order. For large Authority Files, a limited hierarchy might be employed to make access easier. See Section 3.3 for more information. **Glossaries** are lists of terms from a subject domain with accompanying definitions. **Dictionaries** usually have more general domain application than glossaries and may include different senses of a word meaning. They are always presented alphabetically and may have information on word origins. **Gazetteers** list place names and may also include coordinate information on locations in various types of 'footprint', such as centroid, bounding box, etc. **Synonym Rings** have recently emerged as a type of term list, offering *synonym control* in (free text) web search tools. They are not used for indexing purposes but give the option of synonym *query expansion* of a concept in free text (uncontrolled) search engines. For example, Google has recently added an option of synonym expansion to searching, while domain specific sets of synonyms can be found in search engines for particular websites.

3.1.2 Taxonomies

All taxonomies provide a hierarchical organization of categories. The hierarchical relationship may be loosely or more specifically defined. They usually serve a *classification purpose* (similar items are grouped into the same "bucket" – see above). As

such, they can be considered as (simpler) examples of classification schemes. Complex examples from the Library domain, such as the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC), are considered below under classification schemes.

Hierarchical organization of information occurs in many domains and various forms of taxonomies exist, serving different purposes and organized by different types of characteristic of division. **Taxonomy** is a particularly loose term, with a wide usage even within terminology circles, varying from relatively simple menu systems to complex corporate knowledge bases. Taxonomy is associated with (at least) three different communities: scientific taxonomic systems, website designers, corporate taxonomies. Examples from the sciences include the well known scientific taxonomies – see the life science projects discussed in Section 5.2.5. In website design, taxonomy is the most common term for a variety of terminology systems, sometimes very informal. Taxonomies are used as the basis for menu systems, as a method of organizing a website to facilitate interactive browsing through sections of the website, or to underpin other access mechanisms. In some situations, a very loose hierarchical relationship is employed to structure the menu system. Sometimes the menu structure is dynamically generated from an underlying knowledge base. In some business information environments, with different tailored views possible, we approach a more general corporate knowledge management structure. Various web development applications attempt to provide some form of automatic creation of taxonomies (see eg ch16, Rosenfeld and Morville). However human input is also recognized and the new role of 'Information Architect' has emerged. Daniels and Busch (2005a, 2005b), from the company Taxonomy Strategies, review commercial use of taxonomies and discuss best and worse practices, also considering vocabularies with regard to particular Dublin Core metadata elements. They recommend factoring the DC Subject element into separate facets when appropriate and give examples. ROI issues are discussed (see also Rosenfeld and Morville, ch 17-18).

To be useful, it is important to remember that more is involved than creating a simple hierarchical structure. Consider an example from a case study of Microsoft's successful application of taxonomies (considered broadly) to the internal MSWEB, (described in detail in ch20, Rosenfeld and Morville). The Microsoft team's use of taxonomy encompassed: hierarchical controlled vocabularies with equivalent terms for the same concept; metadata schema of the attributes for a given document type; category labels for the displayed options in menu systems. Their tools included a Vocabulary Manager (supporting the editing of vocabularies and relationships between them, including thesaurus relationships), a Metadata Registry and a URL Cataloguing Service.

Recommendation: It is important to consider the broader context and return on investment

3.1.3 Subject Headings

Subject headings are controlled lists of subject terms. They often have broad coverage but with shallow hierarchies. They usually allow for 'coordinated', composite headings,

formed by combining single subject terms according to rules. These rules may be more restrictive than a faceted classification. Well known examples include Library of Congress Subject Headings (LCSH) and Medical Subject Headings (MeSH). They typically have a set of main headings which may be allowed to have subdivisions or qualifiers appended.

3.1.4 Relationship-based KOS

Relationship-based KOS are defined in terms of concepts and more clearly distinguish between different kinds of relationships than the previous KOS structures, while varying in granularity of relationships and degree of formality of definition. There tends to be a practical trade off between expressivity (eg number of relationships) and both interoperability (via common agreement on meaning and use of the relationships) and overhead in design. The common KOS variants tend to overlap in structure but are designed with different purposes in mind.

3.1.4.1 Thesauri

The **thesaurus** is designed for retrieval purposes and has a restricted set of relationships. The three thesaurus relationships are Equivalence (connects a concept to terms that act as effective synonyms), Hierarchical (broader / narrower concepts) and Associative (more loosely related, ‘see also’ concepts). These are defined by international standards. The British and US standards have recently been revised and extended (BSI still ongoing, NISO). The standards discuss common subtypes of the three relationships. For example, the hierarchical relationship can be specialized into Generic (subclass/superclass), Instance (class/instance) and partitive (whole-part) relationships. The equivalence relationship connects a concept with a set of equivalent terms, treated as synonyms for the retrieval situations envisaged by the designers, and again various subtypes are possible. Either mono or poly hierarchical structures may be employed.

According to the thesaurus standards, assertion of relationships between concepts is governed by strict rules. Some widely used thesauri do not follow all the rules but still appear to function effectively for their purposes. Thesauri tend to be defined for a particular subject domain or family of products and can be large. They are usually employed for descriptive indexing purposes and corresponding search systems. Thesauri can also be used as a query expansion resource in free text search engines (sometimes then referred to as “search thesauri”).

3.1.4.2 Classification Schemes

Classification Schemes in many ways are similar to Taxonomies (above). The more complex classifications, with well defined hierarchical relationships should be considered as relationship-based KOS. Well structured classification schemes are mono-hierarchical, conforming to the principles of exhaustivity (covering all relevant subjects) and mutual exclusivity. Complex schemes, such as DDC, augment a concept with a wide range of auxiliary information and connections, including sets of effective synonyms, ‘see also’ cross-references within the scheme, direct and looser (eg co-occurrence) mappings to concepts in related schemes or thesauri, etc.

There are two approaches when dealing with compound subject descriptions that combine individual concepts. In **enumerative schemes**, all legitimate combinations are explicitly specified in the scheme and located at a place in the class structure. Any new compound subject must be explicitly added to a new version of the scheme. Alternatively, there may be rules to express valid combinations (synthesis rules) by combining atomic concepts and this allows a much wider range of subjects to be described than is practical to explicitly enumerate. Such schemes are called **synthetic**. In practice, there are also hybrid approaches.

3.1.4.2.1 Faceted Classification Schemes

Faceted systems apply facet analysis to the process of synthesizing complex descriptions from atomic elements. The term, facet, is used in different ways which gives rise to some confusion. In this context, it normally refers to a set of fundamental categories (as appropriate to an application domain) and their combination according to rules. Each fundamental category might itself be a class hierarchy. Most commonly the different facet dimensions are mutually exclusive. Single concepts from different facets are combined together when indexing an object - or forming a query. Often this is a simpler and more logical organization than attempting to form a single hierarchy that encompasses all different possible combinations of (e.g.) objects and materials and agents.

Faceted browsing interfaces to web databases are useful when a user is able to orient to the initial display and various commercial search engines now offer this facility. Pollitt's HIBROWSE system demonstrated the potential of browsing facet hierarchies and interactively combining terms from several facets to refine a query (Pollitt, 1997). The Flamenco system dynamically generates previews of query results as the user browses different facets (Hearst et al. 2002; Yee, 2003). Some user evaluation has been conducted and Flamenco is now available on an open source basis. In the UK, the Adiuri faceted system has been used to develop Web interfaces to some JISC Projects (see the Common Information Environment (CIE) Demonstrators in Section 5.1.5). Faceted, 'filter-flow' interfaces can guide the user through a set of choices, dynamically updating the range of options with each choice. Faceted approaches to searching may also be helpful in situations where query rather than browsing is appropriate (e.g. deep/unfamiliar hierarchies) or when query preview is impractical (Tudhope *et al*. 2006).

A somewhat simpler notion of facet is prevalent in many of the Web interface contexts and in the USA. Here facets are often different metadata elements and there is little notion of the semantics of combining them (see for example NISO and Rosenfeld and Morville 2003). Facets might include Place, Time, Price, Colour, Audience, etc. and may not always be hierarchically organised. See also the XFML representation for a class of faceted web interfaces in Section 6.2.

In the UK, influenced by the work of the Classification Research Group, more complex faceted systems can be found. Facet analysis is applied to different aspects of Subject, all hierarchically organised. Here fundamental categories might include Abstract entities, Objects (of different types), Materials, Agents, Processes, etc. Different types of rules

govern the ‘syntactical’ combination of facets (sometimes recursively at lower levels) and an ordering principle is often applied, useful for structured browsing. For more information, see BSI Part 3; Aitchison *et al.* (2000).

Faceted Classification Schemes are similar in some ways to faceted thesauri, such as the Getty Art and Architecture Thesaurus, and to some ontologies (eg Section 5.2.6 - OpenGalen).

Recommendation: Consider faceted approaches when developing vocabularies and TS

3.1.4.3 Lexical Databases

The most well known lexical database is Princeton University’s WordNet, which is a general purpose linguistic resource, with a wider range of semantic relationships than thesauri. There are separate databases for nouns, verbs and adjective/adverbs, each with its own set of relations, including hierarchical relationships. WordNet distinguishes between different word senses via domain-independent lexical relationships, including homonymy, antonymy and synonymy (extensive “synsets” are provided). It has been employed in a wide variety of general language processing applications, although other lexical databases might well be used for specialised purposes. An EC Telematics Project produced a EuroWordNet, with different European language versions.

3.1.4.4 Ontologies

The term, **ontology**, is sometimes used loosely for any knowledge organization system, particularly if it is represented using Semantic Web standards, such as RDF. However as intended for AI modelling and inferencing purposes, ontologies tend to have the most precise and formal definition of relationships of the knowledge systems discussed here. An ontology will contain classes (concepts) and instances of those classes, being objects in the domain. Classes will usually have attributes so that complex objects in the domain can be described. Relationships will include is-a (for class hierarchies), instance, partitive and (sometimes many) domain specific relationships.

A distinction is made between detailed **domain ontologies** (which can be thesauri or classification schemes or enriched versions) and more general **upper (foundational) ontologies,** which describe fundamental rules and axioms governing relationships and their composition. **Core ontologies** seek to act as unifying frameworks for a general domain, sometimes bridging different domain ontologies. The CIDOC Conceptual Reference Model (CIDOC CRM) is a widely used example from the cultural heritage domain. Ontologies can be associated with formally defined axioms and rules for processing and combining relationships and are intended for use with logical reasoning systems. Consequently, they are suited to applications with well defined objects and operations (see Section 3.2.3).

3.2 Vocabularies by purpose

We now consider different broad purposes, communities of practice and intended contexts of use (allowing for some overlap in practice). Folksonomy and name authority services are considered separately due to current topical interest.

Recommendation: Descriptions of intended purposes of a vocabulary would be a useful element of a vocabulary registry (see also Recommendation 3.7)

3.2.1 Retrieval purposes

Information retrieval KOS are intended primarily to assist retrieval of resources, originally from bibliographic databases and library catalogues and now from Digital Libraries and the Web. The design rationale is perceived assistance in future retrieval operations. These include classification and indexing, search (including browsing, query and various forms of "intelligent" searching), mapping between KOS (mono and multi-lingual), providing a framework for learning a subject domain or exploring it in order to refine a (re)search question (defining concepts and setting them in context). A KOS might be used both for classification/indexing and searching, or just searching. KOS can be used to support manual cataloguing and also automatic cataloguing activities. KOS range from domain specific KOS to general classification systems, from two hierarchical levels to systems with great depth and breadth of coverage.

3.2.1.1 Classification vs Indexing

The distinction between ***classification*** and ***indexing*** is important but often misunderstood, especially in new Web developments (Lancaster 2003 is a good text). Both processes assign descriptors or tags to information resources. Both can involve KOS with hierarchical arrangement of concepts. However, classification seeks to group similar items together, whereas indexing seeks to bring out the differences between items, in order to help distinguish them during search. Classification provides an overview and assists organization of material. This structure facilitates methods of access based on browsing, whether browsing library shelves or hierarchical menu systems. **Classification Schemes** are often associated with a notation or coding scheme that produces an ordering, useful both in shelving and in ranking results of a search. Indexing (eg with a **thesaurus**) seeks to be more descriptive of an item's content, as opposed to assigning an item to a broad category. Thesaurus descriptors may be combined during search. The difference is sometimes compared to a table of contents versus a back of book index. While the structure of a classification system and a thesaurus may be fairly similar, in that both consist of hierarchical structures of concepts, they will tend to differ in the *exhaustivity* and *specificity* of their application to information items. Thus an information item will generally tend to be classified by fewer, more general concepts from a classification system and conversely will tend to be indexed by several, more specific concepts from a thesaurus.

Sometimes a classification and an indexing system are combined to cover both purposes, for example a classification scheme with a thesaurus. This affords much flexibility in browsing interfaces and rich resources for automatic classification and search tools. It can

also be very useful in offering different classification-based filters on (thesaurus-based) search results.

3.2.2 Linguistic purposes

KOS are used as resources for various natural language text processing techniques (both automatic and intellectual), including the areas of machine assisted translation and language engineering with named entity extraction, text mining, summarization. Section 6.5 briefly points to some language technology standards. The term, 'terminology' is often used by the natural language community to refer to a language-purposed vocabulary.

Commonly used general purpose, linguistic resources will tend to contain finer grained relationships for language engineering purposes and are briefly discussed under **lexical databases** (Section 3.1.4.3). However some information retrieval KOS can be considered as containing elements of linguistic resources, for pragmatic application of natural language techniques. Thus, significant thesauri and classification systems will have a large entry vocabulary of terms considered equivalent for the envisaged use contexts. They may contain extensive scope notes or definitions of different kinds, which can be viewed as linguistic resources.

3.2.3 AI purposes - modeling the entities in a domain

The term, 'ontology', derives from metaphysics, a branch of philosophy concerned with the description of reality (Smith 2003). It was adopted by the AI knowledge representation community, although the AI use has some differences. An ontology

> … is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.
> …
> Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. We build agents that commit to ontologies. We design ontologies so we can share knowledge with and among these agents.
> …
> A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly.
> (Gruber)

As outlined in Section 3.1.4.4, AI ontologies are formal representations, modelling a knowledge domain with precise definitions and relationships. They are designed to be used by first order logic reasoning systems and are a knowledge representation mechanism for communication between (automatic) intelligent agents. They are often associated with Semantic Web research and database schema integration. They are suited to applications with well-defined objects and operations, situations where it is possible to reach agreement as to the precise definition of concepts (and terms) and where it is useful

to define logical rules for processing relationships and possibly inferring new knowledge. These applications tend to have a different focus than retrieval per se, for example elements of analysis in eScience, or automatic generation of new data. Examples might include many scientific applications, where the ontology is a model of currently accepted scientific knowledge and smaller subject domains, such as some business applications. There is overhead in creating (and sustaining) formal representations and in some situations it may not be not feasible to come to commonly agreed, precise definitions on abstract or contested concepts (eg some descriptions of human activity). For example, in search applications, where a fuzzy notion of 'aboutness' is the basis for indexing or classifying a document, as opposed to an assertion of fact, a less formal approach may be suited.

3.2.4 eLearning purposes

The field of eLearning covers a variety of applications and projects, including work on eLearning Repositories, VLEs and projects with dedicated material. To some extent, vocabulary work in the eLearning and Digital Library fields tends to take place independently and increased cross-fertilisation would be a beneficial future development. Collaborative examples include the ADEPT Project, which investigated the combination of structured vocabularies and visualisation techniques to assist students learn scientific concepts in an applied context (Smith *et al.* 2004) and the recent RDN/LTSN partnership (Powell and Barker 2004).

Recommendation: Increased cross-fertilisation between eLearning and Digital Library fields

The eLearning domain has seen an emphasis on standards for Learning Object metadata, with various elements recommended to be taken from relevant controlled vocabularies. The IEEE Learning Object Metadata/IMSL Learning Resource Metadata is a standard, which allows comprehensive description of the different aspects of a learning resource. The JISC Centre for Educational Technology Interoperability Standards (**CETIS**) has useful sets of Metadata standards briefings, along with links to standards and guidelines.

Like other domains, eLearning applications attempt to make use of terminology for more effective cataloguing, sharing, discovering and retrieving objects in the domain. However, eLearning has some distinctive aspects, due to the specific nature of the *learning* resources that are its particular focus. **Learning Objects** (LOs) are complex entities that can be accessed in different ways and that combine different use perspectives – for example learner(s), teacher, developer, digital librarian. The granularity of learning resources might vary from a complete course to a very fine grained LO, with one precise objective. They may involve multimedia elements and may involve design and control of navigation paths. There is interest in mechanisms that might allow primitive LOs to be defined and combined together in instructional sequences.

Potential access points for LOs subsume pedagogical dimensions, in addition to the Library's traditional subject or topic based description. The different dimensions might not be considered orthogonal; for example, in some situations, appropriate subject

description terminology might vary with the intended educational level of a learning resource. Even when considering topical subject description independently, it can be approached from the point of view of a general subject discipline or from a curriculum related perspective. Thus the LO selection process, employed by a student or teacher or indeed indexer, may involve multi-faceted relevance judgements in ascertaining whether a particular LO (perhaps the result of a search) is appropriate, as opposed to bibliographic topical relevance judgements. Knowledge of how people use LO vocabulary elements in practice would be useful.

Recommendation: User studies of behaviour by indexers (cataloguers), students, teachers. Investigate how to support effective practice with a variety of indexing and retrieval tools

In eLearning applications, vocabularies tend to be **taxonomies (and classification schemes)** or **term lists**, although **thesauri** are also found, particularly for subject description. Within eLearning, the IMS Vocabulary Definition Exchange (**VDEX**) is an open specification for the representation and exchange of the eLearning vocabulary types mentioned above. Faceted and poly-hierarchical vocabularies can be represented. An Information Model and Guides for Best Practice, Implementation and an XML Binding are available (Fegen 2006). The XML binding allows application of XML style sheets for tailored views. However an RDF binding is not available.

Recommendation: Investigate conversion between VDEX and SKOS Core representations for compatible vocabularies (see also Recommendation 6.2)

The full IMS LOM data model consists of 9 basic categories: General, Life cycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation and Classification. See Barker (2005) for an overview and the IMS Meta-data Best Practice Guide for an extensive description. **Vocabulary** type elements are constrained to be drawn from a specified controlled vocabulary and are Source-Value pairs. The **Classification** category (which can be repeated) is probably the most relevant for general vocabulary issues, although vocabulary based elements may occur in the other 8 categories, particularly examples of term lists. The Classification element is complex with various sub-elements, allowing description of different Purposes (or perspectives) for classification and, via the **Taxon Path**, source and taxonomic identification. Free text descriptions and keywords are also possible to complement controlled vocabulary or allow more specific descriptors for particular applications.

Indexing might draw on various aspects of the **Purpose** sub-element of Classification. Some tend to be free text, such as **accessibility restrictions; prerequisite** requirement; **skill level; security level; competency.** Vocabulary controlled elements often include:

- overall subject **discipline** (e.g. CanCore recommends a simplified version of DDC)
- **idea** or concept (for example a particular discipline's vocabulary such as AAT in the arts, ERIC or BET in education, MeSH in medicine, or a curriculum based vocabulary such as JACS)

- **educational objective** which may be free text, though some profiles have recommended vocabularies (e.g. RDN/LTSN LOM recommends LTSN pedagogic terms and TOIA-COLA draws on Bloom's Taxonomy)
- **educational level** when an LO has a specific educational target audience (eg UK LOM Core recommends the UK Educational Levels vocabulary)

A few indicative examples of educational vocabularies are listed in the eLearning references. For a comprehensive listing of educationally oriented vocabularies, see Report 1 from the **JISC Pedagogical Vocabularies Project**, which also has more details on eLearning vocabularies and related projects in general. This includes the **Becta Vocabularies Studio** (hosted by the Vocabulary Management Group), which supports editing and maintenance of vocabularies (see also the JISC Shared Infrastructure Services Review). There is also a Vocabulary Bank, a repository for educational vocabularies, with a web services interface and a Tagging Tool. The Vocabulary Studio maintains a central spine, used for dynamically mapping between vocabularies. Vocabularies are represented in the Zthes XML DTD. Vocabulary management software is available from the Vocabulary Management Group built on SchemaLogic's SchemaServer engine with open source additions. Basic browsing and searching of the vocabularies is supported.

Various eLearning tagging tools have been developed (see references) and useful eLearning oriented cataloguing guidelines are available from JORUM and LearnDirect. Currier *et al.* (2004) see a continued need for tool development to support both cataloguing and search and for guidelines to support effective use of eLearning vocabularies. They describe examples of projects which experienced difficulty in cataloguing LOs and recommend collaborative teams with expertise in subject, pedagogic, metadata and discovery areas. The consistency problems they describe echo studies of indexing/classification practice over the years, showing low intra and inter – indexer consistency. (See Section 4.1 on studies of user information seeking behaviour.)

3.2.5 eScience purposes

Vocabularies for eScience share the general points relating to vocabularies outlined above. They also retain some of their own particular characteristics. A brief selection is mentioned to introduce some issues.

One major feature is coverage of non-textual material as a basic information element. Thus vocabularies exist whose purpose is to describe numerical datasets. These range from controlled term lists and Authority Files to the more structured relationship-based vocabularies. As with eLearning, they may attempt to deal with very fine-grained data elements and may involve non-topical vocabularies, for example physical units or parameter files for experiments, as well as various types of authority name. In some disciplines, there is a move to specialised markup languages for this purpose. For example, in Chemistry there have been moves to link chemical names to molecular structures and to describe experiments in a structured way, for purposes of re-use. Ontologies have been used recently in UK Grid projects (for example MyGrid).

There is a long tradition of making use of taxonomies and various initiatives facilitating Web-based taxonomic resources have made progress in recent years. Some life science projects are briefly reviewed in Section 5.2.5, including NCBI's Life Sciences Search Engine and Taxonomy Browser. Much effort has gone into vocabulary-based indexing (and searching) the medical research literature and initiatives such as the UMLS metathesaurus have sought to unify different medical vocabularies. Effort has also gone into indexing abstracts or resources with multiple vocabularies.

Recommendation: Studies of user practice with vocabularies describing research data.

3.3 Named entity authority and disambiguation services

Factual data, in the form of named entity authorities, is an important aspect of terminology services. The main function is to identify and use correctly named entities, (a) improving precision and recall in retrieval by joining different name variants of an identical entity and (b) disambiguating identical name forms that refer to different entities. These are the same general controlled vocabulary problems outlined in Section 1.2.3 but they are intensified with name authorities due to the frequency and importance of their occurrence. These problems can be extensive in a single database or repository. They multiply, however, when using different sources for searching or when building aggregator services. Areas of application include support for indexing, linking, searching, browsing, disambiguation, metadata enhancement and terminology creation. Project Perseus (Crane and Jones 2006) found that about 6-7% of all words in text are named entities, i.e. person and organisational names, places, times and dates.

Semantic interoperability efforts have aimed to foster consistency by standardising with the help of, primarily, name authority databases and gazetteers or other geographic name authorities. Text and data mining techniques can be instrumental as a support for such authority files and their creation and maintenance or even as an alternative in some of the application areas.

In more detail, the results of such efforts are needed to

a) support keyword assignment and named entity indexing
b) allow and improve automatic indexing of content
c) support advanced searching and browsing
d) allow metadata validation and enhancement operations
e) allow cross-searching/browsing and linking between several information sources
f) identify potential candidate terms for the creation of a suitable and topical domain terminology and to contribute to the building of domain-specific authority files

3.3.1 Name Authority databases

Libraries, especially National Libraries, have a long history of activities, controlling names and creating name authorities. This was originally aimed at authors in the traditional printed publication world, via printed and online catalogues and national bibliographies. In its most advanced form, this lists all known name forms; identifies a preferred form; provides additional biographical and affiliation information, including

sources to assist in uniquely identifying an author. Each record carries a local identifier number, which can be used to associate records in literature databases with a unique person. Clearly, this level of authority control is quite expensive. The key part of uniquely identifying an author needs to be carried out by humans, even though there can be a high level of machine assistance.

The most well-known effort of this kind is the Library of Congress Name Authority File (LCNAF). Name authority records in MARC format can be downloaded free of charge for use in a local library system. In the UK, the British Library (BL) Name Authority List is no longer used by the British Library. Since 1997 the BL has been contributing new personal name headings to LC NAF and a retrospective merging of the files is ongoing.

On an international level, several European projects support development and integration of name authority records, emanating primarily from national libraries, i.e. LEAF project - Linking and Exploring Authority Files (LEAF).

A national effort in the Netherlands actively integrating academic authors publishing on the Internet with names from the Union Catalogue is the "National Author Thesaurus" (strictly not a thesaurus). OCLC Pica has been commissioned to develop this for the Dutch national digital academic repository network (DARE). (From presentations by Leo Waaijers) 50% of Dutch authors are covered by the National Union Catalogue and another 40% are expected to be added, via matching with the institutional research registration system, Metis. One reason for the anticipated high coverage is that authors of academic journal articles are traditionally not covered by the libraries/National Union Catalogue. Manual additions are expected to lead to a full coverage. Final release of the database is expected for the end of 2006. DARE is actively looking for international cooperation. The ePrints UK project aimed to apply name authority to authors names in the descriptive metadata records the service harvested from UK institutional repositories, but could not find an appropriate source to build upon the requirements of the service.

There are various discipline-specific and organisational name lists available, however the requirements of services differ depending on their content.

The key issue with name authority files is generating the initial data to populate them. Archivists have always recorded more detail than libraries in name authority files, finding this necessary in order to distinguish between names. The National Register of Archives (NRA) has some 180,000 standardised corporate, personal and family names, each of which needs to be developed from the current skeleton record into a full record by the addition of content and links. There are potentially many thousands more, including some on A2A (Access to Archives). Developing the name entries in the index into full authority records is a labour-intensive process, and has so far proved an insuperable barrier to the NRA indexes being launched formally as name authority files. Funding is unlikely to be available within The National Archives (TNA) in the foreseeable future.

In order to progress development, TNA is keen to collaborate with JISC and other interested bodies. TNA is willing to provide leadership and technical expertise to support

the initiative. Preliminary discussions have already taken place with a range of organisations including JISC, The Arts and Humanities Research Council, The Arts and Humanities Data Service, MLA, The Heritage Lottery Fund. It is also planned to include the British Library in discussions.

Subject-specific authority files include the ERIC Identifier Authority List, relating to education, and the American Institute of Physics Authority Database, relating to physics and allied sciences. Professional societies such as IUPAC and IUCr in Chemistry and Crystallography maintain lists associated to their World Directories of researchers and members. Person and institutional names could possibly be extracted from universities and research funding agencies (as in the Netherlands), provided integrity legislation does not prevent such re-use. Commercial enterprises are building services on top of personal and institutional name lists, e.g. the CSA owned Community of Science (COS), claims to have registered about 500,000 researchers from 1600 institutions to assist identifying people with specific areas of expertise.

Reference sources are also authorities. Useful sources for building authority systems include rich, traditional sources such as encyclopaedias and dictionaries, but also the recent, participatory and open encyclopaedia on the web, Wikipedia. This provides authority information about people and organisations, performs name disambiguation, synonym control etc. (Wikipedia). The German National Library has cooperated with German Wikipedia from 2005 in the usage of the name authority files, Personennamendatei, PND. 20,000 out of 100,000 biographical articles in German Wikipedia carry PND numbers (identifiers), which can be used for bi-directional links between Wikipedia and authority records, or bibliographic information about publications in library OPACs (Voss 2005). Project Perseus has also been making use of Wikipedia, finding high levels of correctness. In any case, such reference sources are highly valuable as training data for named entity recognition and text mining purposes.

Compared with the authority systems created by libraries, usually the lists developed by disciplines and organizations (in their raw form at least) are not authority files ‘proper’. Since they do not contain sufficient and unique information, they are not very well suited for, say, disambiguating names. Different name variants may still appear, for example because the association between a name and an organisation is often temporary and organisations can be renamed, split up, merged etc. Modern authority and access systems could assist in the necessary upgrade to unique identifiers for people and organizations in the non-library lists.

In this context, the benefits of standard formats for authority information becomes obvious. The Library of Congress Name Authority Format structures rich information, as does ISAAR(CPF) - the International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (Second edition, 2004) - in the archives world. This includes biographical and historical information about persons/organisations, potentially highly useful for disambiguation purposes. Another standard is the National Council on Archives format. When cross-searching information or joining sources, the lack of interchange mechanisms between different standards is a serious problem. Using

authority databases is an expensive but efficient approach, which greatly improves retrieval performance for users and avoids duplication of details about names in many databases, resources and collections.

With regard to (unique) identifiers for names, name authority lists normally use internal record numbers to identify author names. For authority services that establish the unique person, this identifier becomes a unique identifier for the person, e.g. the LC control number in OCLC's service. Nevertheless, other authority lists and services will possibly have assigned another unique identifier to the same person, severely hampering cross-search.

Unique identifiers should be based on proper authority records, not just on different name forms found. Another requirement for these ID's to be useful is an organised cooperation between service providers running local databases in order to correctly merge or link records for identical persons. Correct name disambiguation requires a proper authority system.

National coordination efforts have potential to be recognised as authorities for people clearly belonging to the country and their identifiers might be widely reused. DARE in the Netherlands runs the "Digital Author Identification" project, carried out in Groningen as part of the DARE "National Author Thesaurus" effort to investigate and solve these problems on national level. In the UK, the Eprints UK project intended to establish authority control. In that context, there was a suggestion that HESA (Higher Education Statistics Agency) identifiers (HEI Identifiers for institutions, work on student and staff IDs) might be a building block, perhaps through a national HESA registry.

At an international standards development level, the IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR) aims to study the feasibility of an International Standard Authority Data Number (IFLA).

Recommendations:
Investigate lists of institutional names and academic affiliations (IESR Agents etc.)
Study the coverage of available name authorities in OPACs and academic web publishing (LEAF, CiteSeer and similar)
Engage in international cooperation (OCLC, DARE, LEAF)
Prototype a demonstrator UK Name Authority File, maybe together with BL and universities (authentication, staff, institution databases) and evaluate its use in a limited application

3.3.2 Other named entity authorities

Apart from personal and organisational names, there are other named entities of relevance to information services for research, learning, memory institutions and Digital Libraries: i.e. place and other geographical names, street names and addresses, dates and periods, products and names for sources or objects of study. The usage of these authorities and their creation is very similar to the ones discussed regarding personal names. Thus, only a few pointers are provided here.

Place and geographical names have attracted most interest and activity. Quite a lot of geo-referenced information is available in digital form. Although not the focus of this review, localised searching is a prominent Internet search feature with heavy investment from major companies in the business. This probably tends to be dominated by GIS coordinate data and map-based navigation, rather than place names. This is also the case with HEIRPORT (HEIRPORT), the portal provided by the Historic Environment Resources Network, which makes available data drawn from different public bodies about archaeology and the historic environment, in collaboration with ADS, the Archaeology Data Service (see also Section 5.1.1). A map interface translates a search for a location or area into coordinates, which are used to search the database. To what a degree placenames are used, mapped to coordinates or standardised upon is not immediately clear.

In the digital library context, two databases and efforts have been pioneering and still dominate when it comes to global coverage: The database of the Alexandria Digital Library (ADL) and the Getty Thesaurus of Geographic Names (TGN). A good overview can be found in the special DLIB issue on Georeferencing (Hill 2004). Many countries have national and regional Gazetteers, as valuable sources of place names. Historical place names pose challenges. Historical placenames in Britain are addressed e.g. by Southall's Great Britain Historic GIS Project (GBHGIS), the historic boundaries being accessible via Edina (UKBORDERS) and placenames and landmarks from three gazetteers from late 19th century can be queried in BBC Vision of Britain through Time (Vision).

JISC is funding two important projects in this field: GeoXWalk (GeoCrossWalk) and the recently started GRADE project (GRADE), investigating the reuse of geospatial data.

Recommendation: Address the treatment of place and geographical names in UK services and activities, and the development of standards and authorities, in cooperation between these projects and related terminology efforts.

Regarding time periods, a recent important project to establish an authority file (or directory, as the project calls it) is the Time Period Directory at the University of California Berkeley, listing named time periods in connection with associated locations (Petras). Interestingly, the initial set of named periods is extracted from a traditional library subject heading system, the Library of Congress Subject Headings, and they try to model a standard after the ADL Gazetteer Content Standard (which operates on place names). The level of integration between these named entity approaches is obvious from the projects further plans: "... development of prototype systems to support the dynamic interaction of Time Period Directories, digital gazetteers, biographical data and ontological structures like thesauri and classification schemes, in combination with a variety of network-accessible digital library resources ranging from library catalogs to archival collections and digitized versions of historical primary resources." The crucial role of topical terminologies is obvious as well as the rich dynamic options of interaction between maps, biographies, timelines and chronologies with primary research materials.

The previously mentioned Perseus project (Perseus) has broad activities with regard to named entities. It has explored text mining techniques for creating different authority lists for predominantly historical texts and has developed useful visualization and navigation options based on this. While the lists themselves might not directly be useful for very different geographical, historical and cultural contexts, the approaches and techniques almost certainly are (for an overview, see Crane 2004).

For each geographical and historical/cultural context, usually separate efforts for creation of authority databases are required. They might later add up to broader coverage via international and inter-disciplinary cooperation, if this is deemed useful. However, in the natural and biological sciences particularly, named entities have to be managed and standardised by international bodies and professional organisations. This is happening to a varying degree and many efforts do not necessarily reach the level of established and used standards. In Chemistry, e.g., there are International Chemical Identifiers, Chemical Formula and IUPAC Chemical Names and so-called Colour Books for other terminology at a varying degree of standardisation and alongside commercial "standards" such as the ones from Chemical Abstracts.

Recommendation: Support active participation of UK institutions in international naming standardisation efforts in scientific disciplines and, via project support, assist their implementation in UK

3.3.3 Named entity recognition, text mining, name disambiguation

Named entity recognition (e.g. via text mining methods) cannot fully replace authority systems, since only identifies text strings representing named entities, it cannot finally decide which name forms are correct and which are fully equivalent to a given unique named entity. From the context a name appears in, certain assumptions can be made with different degrees of probability, via approaches such as co-reference resolution to identify variants of names for the same object. However in cooperation with existing authority files, the performance of these techniques can be greatly improved. Vice versa, authority files and services can be expanded and improved, based upon text mining from publications and other suitable sources.

Name disambiguation services work best based upon appropriate and comprehensive authority services. In narrow disciplines and more coherent collections, named entity recognition and co-reference resolution may be capable of providing a sufficient level of disambiguation, depending on ambitions for the quality of service and a thorough cost-benefit analysis. All approaches covered in this section essentially assist name disambiguation to a greater or lesser degree, but do not necessarily provide a complete solution.

In a broader context, text and data mining can be used to improve both repositories and aggregator services, with respect to subject access and terminology use (see section 4.6). There are various possible purposes for applying mining and the specific techniques needed and the approaches relevant will accordingly be different. Particular techniques

will be used in relation to named entities. Text mining can be applied in order to identify, link, search, browse, index and extract named entities, such as author and institutional names or e.g. chemical compounds and their components, via identifying more or less standardised named entities, or their substrings from textual materials and publications. Probably equally important is the extraction of relevant topical terms and phrases. Apart from repository documents, other collections such as corpora of text and data and reference works are needed.

When using data mining with e-science databases and data (centric) repositories, additional features might be needed for knowledge extraction and hypothesis creation, data selection and comparison, correlations, calculations etc. Not all disciplines and sources are equally suited to mining and knowledge extraction. Lynch (2006) points out, that certain disciplines may be in a favourable position for early and successful exploration of such options and highlights the role of terminologies: "Areas such as biomedicine or chemistry, where much of the literature is relatively well-structured and where a base of investment in the ontologies, specialized vocabularies and vocabulary mappings and similar tools has been extensive, would likely be fertile ground for early advances." Lynch points out the important role of incorporation of markup to facilitate computational processing.

There is a rich literature on text and data mining methods and techniques (eg Witten and Frank 2000). The next section outlines some projects and tools.

Recommendation: Apply methods of name extraction and investigate their benefits compared to and in combination with traditional authority systems. Build and evaluate different name disambiguation demonstrators.

3.3.4 Tools, Web services

The Library of Congress, National Libraries, LEAF project partners and OCLC ran prominent projects and have tools for creating and maintaining the traditional name authority databases. OCLC runs VIAF, the Virtual International Authority File project. Using its software for matching and linking authority records for personal names, authority records from Die Deutsche Bibliothek are matched to the corresponding authority records from the Library of Congress. Shared OAI servers will maintain the authority files and provide user access to the files (VIAF). The combined approach involving national data about researchers from publication repositories is represented by OCLC Pica who carries out this task for DARE.

As a value-adding service OCLC Research developed a Name Authority control (interactive and automated name authority look-up service and web access to authority records in its Linked Authority File), originally for the ePrints UK project (LC Name), as metadata creation support to be hooked up with templates in the DSpace repository software package. Created as web services this and other developments of the Metadata Switch project/Terminology services project at OCLC (OCLC Metadata) can provide remote semantic interoperability enhancing functionality pluggable into local applications.

The Perseus Project at Tufts University has a longterm specialisation in named entity recognition/mining resulting in quite good levels of results (Smith 2002). They are planning an open source release of the software tools and to offer a service (Crane and Jones 2006). The GATE project (GATE) has developed tools for automatic tagging of personal names (the technique is now integrated into the Greenstone DL software as well). The University of Sheffield, where GATE is developed, plans to use text mining in digital 18th Century materials (Armadillo).

Elsevier's Scopus Author Identifier aims to automatically match name variations and to disambiguate between similar names (STLQ). CiteSeer has made serious efforts in name correction (e.g. with user participation) and name disambiguation using clustering methods based on naïve Bayes and SVM models (Han et al. 2005).

The US project NORA and the University of Illinois, Urbana-Champaign, use a tool for rapid flexible mining and machine learning, including a visualization tool, Data to Knowledge (D2K), which is available on an academic and research license. For other work relating to visualization tools, in the context of text and data mining, see (Fayyad *et al.* 2001; Shneiderman 2002).

Cornell University leads developments regarding metadata enhancement tools and services for the National Science Digital Library (NSDL) project, together with partners such as INFOMINE (University of California, Riverside) and its iVia Virtual Library software. Metadata augmentation, apart from enriching metadata records with subject headings and keywords (subject authorities), can comprise transformation services to correct degraded terms from controlled vocabularies and recognize values from recommended vocabularies, ascribing the appropriate vocabulary encoding scheme to statements. New metadata values can be generated, based on mappings between schemas or vocabularies (Hillmann *et al.* 2004).

Guidelines published in an article on the improvement of metadata quality in ePrint archives (Guy *et al.* 2004 - in the context of the ePrints UK project) underline the importance of early decisions on the usage and granularity of controlled vocabularies, their consistent application and the importance of built-in support for them in metadata editing tools.

Acknowledgement: For the material relating to archives and several other valuable pointers, thanks to the authors of the JISC Infrastructure Shared Services Review, A. Chapman and R. Russell, UKOLN.

Recommendations:
Experiment with a Name Authority Web Service e.g. to be built into metadata creation tools.
Develop or support metadata enhancement services for correction and enrichment: vocabularies, schemes, mapping, names.

3.4 Social tagging and folksonomies

The enthusiastic publicity regarding social tagging and folksonomies (and the broader perspective of Web 2.0) is reminiscent of previous enthusiasm surrounding Semantic Web visions and early metadata initiatives. A balanced approach is needed, acknowledging the value of previous vocabulary work, whilst not ignoring new possibilities.

In the context of this report and upcoming JISC initiatives, it is necessary to investigate to what degree social tagging and similar features have the potential to contribute to an improvement of subject indexing and knowledge organisation and subsequently to benefit resource discovery, browsing and searching. This relates both to possible new and already existing services. Another aspect is the role social tagging and folksonomies could play in creating, upgrading and maintaining vocabularies. Our focus in this review is on the use of social tagging and folksonomies as a contribution to knowledge organisation and discovery, rather than on other potential aspects of social tagging, such as social communication, group creation, bilateral recommendation and personal recommendation lists.

It is important to be aware that many aspects, both benefits and shortcomings, of social tagging are similar to activities known for a much longer time and under different names: author provided keywords (e.g. in scientific articles, Index Medicus and MedLine), user created browsing structures (e.g. in DMOZ and originally in Yahoo), invited user corrections of systems (e.g. CiteSeer) or user-created metadata. In these contexts, user participation concerning keyword indexing, classification and metadata provision has normally not been seen as an undisputed success, with regard to the functionalities of such systems. It remains to be seen to what degree other and new characteristics of social tagging might be more productive, such as the anticipated mass scale of tagging, potential convergence of terminology through public exposure, direct access to most of the sources involved, support through easy to use tools and visualisation, community-based user interfaces (with access to other peoples tags), along with the realisation of private and immediate rewards. These characteristics might, apart from any scale factor, increase speed and reduce costs, offering new qualities in user oriented information services. Clearly, as a public sector infrastructure providing institution, the JISC has a potential role to play here in terms of encouraging the development of trustworthy, sustainable and freely available services.

3.4.1 Terminology

A wide variety of terms are used to name the participatory activity which is the focus of this section, many of them rather mis-leading. Folksonomy (see Wikipedia entry) is often used as a synonym for ‘social tagging’, rather than identifying the whole of the vocabulary space emanating from tagging activity in a specific service. In this report we use the term ‘social tagging’ (although Wikipedia alone gives 16 different meanings of tag and tagging). In the following sections, we go on to delineate various, specific features.

3.4.2 Context

We need to be aware that social tagging exists in the context of a broad range of participatory or community activities in information systems, sometimes a component, sometimes an enabling, overlapping or alternative feature. Related activities include: Linking; Citation; Annotation; Recommendation; Lists (reading, shopping etc.); Exploration of usage popularity, user behaviour and preferences; User contributed metadata; Collaborative filtering, Social searching; Group learning; Customization and personalization (if shared with others).

These activities tend to have in common that they involve communities of use, predominantly secondary and tertiary resources, metadata, opinions, judgments and evaluations, notes and usage experiences, personal views and preferences. Aims include stimulating re-use through reference and recommendation, participating in and contributing to information services, supporting collaboration, cooperative research, learning (and entertainment).

It would be useful to pursue some theoretical effort to systematize and structure the field of activity broadly described as ‘social tagging’, to study purposes, methods and outcomes, as a framework for further research, development and comprehensive suites of features and services.

3.4.3 Categorization of tagging systems

The narrow field of social tagging systems has been categorized by several authors (eg Hammond *et al.* 2005) by: content creator and tag users: oneself or others; by audience: scholarly or general and by object type: web pages/bookmarks/blogs (delicious, Connotea, CiteULike, Technorati), pictures (flickr), music (Last.fm), products (Amazon product tagging), news (Digg) etc. Thus, social tagging is predominantly associated with publications outside traditional channels such as pictures, music, blogs and news. The number of differing systems and applications is so far not very high and the differences rather small.

In our context, services enabling tagging by and for a scholarly audience and covering relevant media are most interesting. Several media types need to be considered, such as traditional primary publications, books, journal articles, museum objects, objects in repositories and data sets, name and organisational directories, terminology systems etc.

3.4.4 Disadvantages and problems

Current publications about social tagging (many in blogs) provide long lists of advantages and disadvantages of social tagging and tend to be written by enthusiastic advocates in a highly promotional vein. Few evaluative, systematic studies from professional circles in knowledge organization, information science or semantic web communities have appeared to date. In order to stimulate future projects and improvements, potential disadvantages and problems of existing social tagging systems are discussed, as regards the scope of this review. Potential benefits are reviewed in the following section.

Compared with traditional knowledge organisation, social tagging redistributes costs, moving them from term assignment to discovery, or as Ian Davis (2005) puts it, "Tagging bulldozes the cost of classification and piles it onto the price of discovery".

An obvious issue with existing social tagging systems is that they are not designed for information discovery and retrieval. They tend to combine various functions in the same application and the same approach is often applied to very different object types: text, web pages, link lists, blogs, pictures and other media, multimedia etc.

The most obvious and often mentioned shortcoming is the lack of any control of the vocabulary. Most harmful to retrieval performance is the lack of simple control (irrespectively of being applied at the time of input or as later improvement processes applied by the system), such as control of word forms (singular, plural), morphological forms (nouns), spelling, use of numbers, character sets and transliteration. A certain linking of synonyms and disambiguation of homonyms is crucial for acceptable recall and precision when searching, as is control of how names are presented (first names, last names, initials, nicknames). Equally crucial are place names, dates/times and acronyms. Also missing are the advanced benefits of KOS, regarding concepts, semantic relationships and controlled mapping of terms.

Another shortcoming is the absence of rules for indexing/tagging: rules concerning exhaustivity, specificity, granularity, compound construction or provision of context. Tags indicating a personal context (e.g. my brother) may not be useful to the public. Place names acting as a subject/topic should be differentiated from personal associations (e.g. a picture of a car photographed in Labrador has not Labrador as subject; it does not say anything meaningful about Labrador, nor about the breed of dogs with that name). The absence of rules e.g. about phrases (often prohibited) or construction of compound terms leads to various ad-hoc practices, with different special characters used for connection of words or other information encoded into the tags, such as creation of hierarchy and structure, or other non-topical metadata (places, coordinates, times, names, types).

The lack of structure among the tags, deprives the systems of concept-based navigational options, such as systematic browsing exploring hierarchies, or other forms of semantic relationships. Alphabetical, indexer-name or popularity sorting are often the only options, hampered in addition by the previously mentioned lack of control. This results in a lack of context for the tagged information items.

In summary, while social tagging may have other benefits, as currently constituted, it is not suited for targeted effective search or systematic topical browsing. In retrieval terms, social tagging systems would have low precision and low recall. The only discovery approach which might be favoured is serendipity. There is, of course, some benefit, in situations where there would otherwise be a complete absence of any other indexing and discovery features.

3.4.5 Advantages and benefits

The major, general benefits with existing social tagging systems include the likely ability to:

- provide tags to many resources where controlled indexing information is lacking
- be topical, evolve and reflect change quickly
- reflect user language without major information loss (helping to preserve language and cultural richness)
- be user-centred
- offer insights into user behaviour.

In addition, there may be potential benefits for ambitious information systems addressing the HE sector in the following contexts:

- indexing input for materials and publications largely ignored by other services, such as images, mixed media and multimedia, learning objects and research data
- smaller co-operating groups, specialised subjects, communication intensive work environments, fields no suitable vocabulary systems exist
- improving entry-level terminology for existing vocabularies, based on availability and processing of user vocabulary
- creating new vocabularies, based on availability and processing of user vocabulary
- improving systems/services by additional data originating from associated social tagging features
- by combination with KOS and subject access features, a different and complementary layer of indexing can be offered
- generally stimulating development and research efforts

Social tagging cannot and should not replace other indexing and knowledge organisation efforts. For the purpose of resource discovery at least, the main recommendation is to explore their strengths and to use them in complementary ways, both by optimising such systems for discovery and by combining them in different ways with more controlled knowledge organisation and retrieval systems. More detailed development and research suggestions are outlined in the two following sections.

3.4.6 Proposed developments

Three directions are proposed to stimulate development efforts and experiments: 1) Improvement of existing social tagging systems; 2) Development of alternative tagging systems; 3) Integration of social tagging features into existing systems and services. The latter two groups of suggestions appear more promising and should be prioritised.

1 Improve existing social tagging systems

- Richer system support should be offered during the tagging process, such as keyword extraction and proposal, dictionary lookup (existing objects, authors, taggers, tags), interactive term disambiguation and visualizations.
- After the initial tagging carried out by users, the system could apply tag improvements (e.g. correct misspelling, specify the language, treat compounds

properly and consistently, link between synonyms, create partial hierarchies, create facets, (see fac.etio.us).
- Search and browse functionalities could be improved via advanced clustering, exploring co-occurrence, other aggregations, filters, ranking and visualization supporting navigation.

2 Build alternative tagging systems

To stimulate innovation, guidance and improved praxis, the creation of alternative tagging systems with different approaches could be supported, such as:

- create a social tagging system optimised for discovery and retrieval
- build a system for a homogeneous service, with well-defined user group and purpose
- explore systematic use of controlled vocabularies in a tagging system
- apply advanced mapping of tags to facets, established KOS or authorities for named entities
- hook library and discovery services into a social tagging system, in the way OCLC does with search engines and institutional repositories (OpenWorldCat in Google, Name Authority service in DSpace).

This development direction has a parallel in the effort of the Nature Publishing Company, to create a tagging system specialised in bibliographic information and reference management based on Connotea (Lund *et al. 2005).*

Recommendations:
Experiment with combination of KOS-based controlled indexing with an established vocabulary and free (social) tagging for research purposes in a specific discipline, optimised for discovery and retrieval
Experiment with potential for automatic linking of tags to facets, controlled vocabularies and authorities

3 Integration of social tagging features into existing systems and services

An important strand of work when it comes to realising the benefits of social tagging is to integrate user contributions into existing information systems and services. This is a special case of the application of all kinds of participatory approaches. In the example list below, we focus quite narrowly on the provision of tags. Addition and integration of different types of user participation to established services might be equally rewarding.

Among the most promising options are:

- OPACs (Online Public Access Catalogues, for library materials): here, some activity can already be seen, e.g. PennPal or the OPACi prototype from Casey Bisson. The immediate activity would be to allow users to add tags to bibliographic records and then using them for different "views", linking etc. Other options would be inspired by Amazon as e.g. the Open WorldCat reviews.
- Subject Gateways, Intute hubs: User tags could be used to inform resource selection for final inclusion and to support improved subject access as in OPACs above. User annotation has been tried before, e.g. in SOSIG Grapevine.

- Directories: Users have been creating and populating categories early on in the Yahoo directory, the dmoz and Open Directory efforts. Yahoo Social Systems has the ambition to do more with the directory.
- Subject repositories: the situation is similar to OPACs and Subject Gateways. There are benefits for resource selection, the acquisition of new indexing vocabulary, the creation of conceptual structures and categories where they are missing now.
- Citation services: CiteSeer already offers the option of user corrections; name disambiguation and adding of content tags would be other alternatives.
- Digital libraries: similar to OPACs, plus creation of structures where there are none. Integration with other participatory efforts would be promising also.
- Search engines: the big commercial search engines are already active in this area, e.g. Yahoo and Google Answer, Yahoo building its whole brand on becoming a "social system". Specialised, academic and local search engines would greatly benefit from similar approaches, integrated with traditional KOS, maybe. Systems like Vivisimo could use social tagging to label clusters.
- KOS creation and development systems might benefit from broader user input and reuse of tags from other systems. The DDC editorial system is opening up for broader expert participation.
- Museum online interactive exhibitions and object catalogues: a few projects are active here, e.g. Steve.museum and the ED2 project at the Cambridge University Museum of Anthropology, mainly with inviting user descriptions of pieces of art and tagging of user experiences (Trant 2006).
- Metadata enhancement services: as with several services above, richness of indexing, disambiguation and correction of errors come to mind.
- Blogs, news services and RSS feeds: these services are natively based on user contributions. More advanced usage of tagging could be imagined, however.

In addition, all these services could, based on social tagging, provide different and alternative views or layers of resources and search results, co-occurrence clustering or automatic linking from tags in the system to external resources such as flickr and delicious.

Recommendation:
Integrate tagging to existing services such as repositories, OPACs, (RDN/Intute) subject gateways, Digital Libraries, KOS creation and management systems, museum exhibitions and catalogues, metadata enhancement services etc.

3.4.7 Research

In the context of social tagging, there are many important aspects which would require research efforts in parallel to service developments. On existing systems now and after the creation of new or improved ones, usage aspects and benefits need to be systematically studied. Very little research seems to be done at this time. Among more specific research topics are:

- study new developments and modifications, compare new approaches to existing ones, look at integration into heritage systems
- develop suitable architectures and study scalability aspects

- study tools and user interfaces and their influence on the practice of tagging, the effects of guided tagging
- look at user behaviour related to tagging and navigation: tagging practice, influence of the social environment and of the display of related/popular tags
- analyse the structure of the emanating "tagspace", analyse the degree of the coverage of the tags in established KOS and the rate of novelty, compare with established terminology creation approaches
- research on the claimed convergence of terminology, mass effects/intelligence (cf. Golder and Huberman 2006)
- investigate the possibilities and benefits of an integration of heterogeneous tagsets
- investigate the benefit of potential tagging standards
- compare with author and professional indexing regarding discovery improvements, study retrieval performance
- look at potential "social benefits" of tagging

Recommendation:
Comparison study between different types of social contributions: annotation, recommendation, personalization, restructuring of information, categorization, concept space, concept maps, topic map tools. This could inform a prototype integrating different types of user participation with social tagging.

3.5 Best practice guidelines for constructing and using vocabularies

Best practice guidelines for design of different kinds of vocabularies offer practical help. Aitchison *et al.* (2000) is a standard reference in the UK for thesaurus design and construction, while the Willpower website offer useful practical guidelines, along with a list of commercial software. As well as describing their respective standards the BSI and NISO standards documents also offer best practice design guidelines, with their scope widening now beyond thesauri. The BSI Guide is perhaps particularly relevant for JISC UK purposes. Middleton's Controlled Vocabulary List includes a Bibliography, list of software and some pointers to Guidelines. The University of British Columbia's Indexing Resources on the WWW contains an extended set of links to guidelines on related issues.

Rosenfeld and Morville (2002) is a widely used textbook for information architecture and website design techniques that build on various vocabularies. It includes chapters on management, ROI and case studies. Daniels and Busch (2005a, 2005b) give specific best (and worst) practice guidelines and ROI considerations from a commercial DC perspective. TASI give an introduction to adopting a vocabulary within a metadata framework. The JISC CO-ODE Project (Section 5.1.2) offers tools and tutorials on ontology design. The GovTalk archive provides design/selection criteria for vocabulary software. The recently published e-Government Metadata Standard (Version 3.1) recommends the Integrated Public Sector Vocabulary (IPSV) as mandatory for its subject element. Concepts from other controlled vocabulary may be added (with the scheme being declared). The IPSV is available in full and abridged versions and in CSV, XML,

RDFS and other formats, while various guides to tagging and use generally are also available.

3.6 Network access to vocabularies

Both for human Web access and m2m access there is a need to discover appropriate terminologies, and to evaluate, navigate and query the terminology once found.

Many terminologies and thesauri are now made available over the Web, intended for human use. However wider use of KOS and their integration into applications in an automated fashion will require m2m access. There are a number of standards emerging adoption of common standards for representing and accessing vocabularies which are outlined in Section 6. There have been some significant steps forward in an attempt to stimulate wider use of existing terminologies. For example, the OCLC Terminology Services project (Section 5.3.6) has recently made available some dozen vocabularies in the MARC 21 Format for Authority Data in XML on its website, in addition to the Dewey Decimal Classification (DDC) 22 Summaries.

There are many web based human readable lists of vocabulary resources (both commercial and freely available), several are detailed in the references. Notable examples include the following. The JISC HILT project has compiled an AtoZ of thesauri. The JISC Pedagogical Vocabularies Project has compiled a list of educational vocabularies. The TASI (the Technical Advisory Service for Images) website has a list of thesauri, classifications and authority lists, along with an introduction to their use within a metadata framework. The Species 2000 website maintains a checklist of online taxonomic databases. The Text Mining Centre has a list of bio-medical ontologies. Middleton's Controlled Vocabulary List includes subject heading lists, thesauri and classification schemes. The University of British Columbia's Indexing Resources on the WWW has lists of classification schemes as well as other vocabularies. Dextre Clarke's extensive review of Taxonomies in the Public Sector also includes sources on potential benefits as well as software and design.

Synapse provides the Taxonomy Warehouse of taxonomies, thesauri, classification schemes and authority files, organized by category. This includes both online links to online vocabularies and their own "value-added fulfillment service" of conversion and packaging with other software. The University of Toronto maintains the Subject Analysis Systems (SAS) Collection, which acts as a "North American Clearinghouse for subject classifications and controlled vocabularies in many different subject areas". The MDA (formerly the Museum Documentation Association) website has a Terminology Bank of cultural heritage vocabularies it has sponsored and publishes online.

In the longer term, it is hoped that such human readable lists might be maintained as 'registries' although the issue of who is to maintain them has to be resolved. See Section 3.7 on registries (and Section 3.2.4 on the Becta Vocabulary Bank).

3.7 Terminology Registries

There are a number of registry initiatives within the education domain designed to support a services oriented approach to component development, providing 'look-up' functionality. These registries provide programmatic access to registered data of various sorts. Such registries include service and collection description registries (e.g. Ockham, JISC IESR), transaction service registries (some using UDDI, mainly within the eScience community, such as GRIMOIRES), registries of mapping and crosswalks (OCLC crosswalks registry - see Metadata Switch Project in Section 3.3.4), and metadata schema registries (JISC IEMSR, The European Library, DART). There has been little activity in relation to terminologies with the exception of the NSDL Schema Registry which plans to register both metadata schemas and KOS related controlled vocabularies in use within NSDL.

In general, registries enable discovery, navigation, access and re-use of the objects that are registered. In relation to terminologies such registries might take different approaches depending on the functionality they are designed to deliver, whether registering descriptions of vocabularies, registering individual terms and concepts, or usage within domains or discipline. Services based on terminologies (such as disambiguation services, query expansion, mapping) might also be registered whether within a specific terminology services registry or within a services registry with a wider inclusion remit such as the JISC IESR.

Providing m2m access to information about terminologies, and terminology services would encourage exploitation of existing vocabularies and enable innovative interfacing with applications from 'other domains'. Policies would need to be established covering status, persistence, identification, and quality.

Registries might be more or less centralised or distributed, depending on policy and finance drivers as well as on technical design decisions. There is scope for co-operation with other international initiatives, both re-using software and exchanging data. To enable interworking and data exchange, registries themselves need to be standards compliant, although standards are immature in this area, particularly around data description. The ISO/IEC 11179 standard has some relevance here (and in particular the XMDR project taking this forward to register more complex structures), however the driving force behind this standards making activity is influenced by database and data dictionary technologies, rather than enhancement of the semantic interoperability of web based services, a focus of more interest to the JISC community.

There needs to be careful consideration of the cost benefit of registries. Whilst some funding organisations such as the JISC might consider registries as a means to identify and promote services available to their communities, there also needs to be investigation of the ROI for providers of specific terminologies and services to contribute to registries. Several of the larger vocabularies have commercial business models and m2m use will raise issues around managing IPR and copyright even for smaller community based vocabularies. There might be a variety of business models, but as with other 'shared

services' it is sometimes unclear who is the obvious funder. Other business issues include clarifying who owns content of a registry? who is responsible for transforming the content of vocabularies to machine readable structure? is there commercial motivation for KOS owners to 'work together' in the context of interoperable registries?

Recommendations:
Demonstrate the use of a terminologies registry within JISC IE testbed to include

- **Investigating inclusion of terminologies into IESR, potentially describing vocabularies as collections**
- **Developing marketing proposition for a UK terminology registry (include use scenarios, IPR issues, business models, cost benefit)**
- **Evaluating use of the draft metadata description profile proposed by NKOS**
- **Maintain collaboration between various UK initiatives (with eScience e.g. GRIMOIRES and learning communities e.g. Becta Vocabulary Tool) and internationally (e.g. NSDL)**

4 Activities with TS

This section attempts to generalise beyond specific projects and types of vocabulary to discuss some ways that terminology services can be applied in wider frameworks. We begin by considering user behaviour, go on to discuss the different types of terminology service in context of the JISC Information Environment and eFramework and finish by considering terminology as part of work in automatic mapping, classification and text mining.

4.1 Studies and models of information seeking behaviour

It is important to consider how people search for information when designing and evaluating TS, in order to reduce the scope for design errors and increase the possibility that services will actually be used. While this is a difficult and complex area, there is a considerable literature on studies of searching behaviour and dedicated conferences, such as Information Seeking in Context (ISIC)[1], have emerged. Such studies offer possible insight into discovery strategies, user needs and user contexts. While remembering that variations in environmental context and individual characteristics, such as training and motivation, can be important, these studies can be a useful resource for planning future developments and evaluation methodology.

The term **information seeking** usually refers to the broader context of an information need, while **information searching** denotes interaction with a computer for a specific search, although the distinction sometimes becomes blurred (Marchionini, 1995; Spink et al., 2002; Wilson, 1999).

[1] ISIC 2006. Information Seeking in Context. http://www.hss.uts.edu.au/isic2006/
Also see First IIiX symposium on Information Interaction. http://www.db.dk/iiix

Terminology support has been found potentially helpful for both recall and precision. In an extensive study of online behaviour by search intermediaries, Fidel's findings (1991) supported the utility of terminology support alongside free text retrieval. According to circumstances both retrieval modes were used to improve either recall or precision. Even professional searchers tended not to make use of synonyms in free text searching, leading to the conclusion that there is a need for well designed thesauri and associated tools.

Information seeking models, such as Choo et al. (2000), Kuhlthau (1991) and Saracevic (1997) provide general frameworks of information seeking behavior which can assist with higher-level design aims. Ellis (1989) critiqued the restrictive assumptions of controlled laboratory evaluations and argued for an empirical, behavioural approach to information seeking studies. This led to focus on basic information seeking patterns, such as browsing, chaining, monitoring, etc. Soergel (1994) stressed the need to take account of the full context of indexing, system and user factors in evaluation. Kuhlthau's (1991) and Marchionini's (1995) models describe the basic stages in the information searching process, in terms of problem definition, query formulation and execution and examination of results. Blocks et al (2006) provide a low level model of the stages of thesaurus assisted search, intended as a practical reference model for system developers.

Studies of searching behaviour generally reveal it to be an iterative process. Bates' influential, 'berry-picking' searching model (1989) emphasised an evolving search, in contrast to models of a static information need where a single query is optimised. She found that in many cases users' information needs evolve as the session progresses in interaction with the material encountered.

This leads to a need to consider the appropriate balance between interactive and automatic TS. For example, the balance between system and user control of terminology supported query expansion (QE) has been the subject of much research. The various Okapi projects conducted a number of experiments with thesaurus based QE in operational settings as part of a probabilistic query model (Beaulieu, 1997). These ranged from fully automatic to interactive QE. Their conclusions favour a balance between automatic and interactive control and explicit versus implicit use of the thesaurus. Other empirical studies considering the user-system balance include Jones et al (1995); Shiri & Revie (2006); Vakkari et al. (2004), and Greenberg (2001), who compared the performance of different thesaurus relationships in automatic versus interactive query expansion. She argues that intelligent systems should take into account (evolving) user retrieval goals.

Research has argued the importance of strategic or conceptual support (e.g. Brajnik et al., 1996; Fidel 1995). Bates (1979) and Fidel (1985) identified a number of *tactics* or *moves* respectively employed by professional searchers to modify or reformulate queries, for example moving to a broader or related term. Bates (1990) discusses possibilities for system support of search activities at different levels of granularity, within a framework of end-user control of the search steps. She argues that one reason current interfaces are difficult to use is that they tend not to be designed around typical search behaviours that promote strategic search goals. She particularly recommends that research be directed to

system support for end-user searching at the mid-level range of tactics and stratagems, as opposed to basic moves and high level strategies.

Recommendation: User studies of TS in context of JISC IE, illuminating the search process (for work flow of services) and the appropriate balance between interactive and automatic TS.

4.2 Information lifecycle with regard to TS

The model presented here draws on the information lifecycle management model described in the DELOS (deliverable D5.3.1, section 3.2.2) state of the art review of semantic interoperability in Digital Libraries (Patel *et al.* 2005), which synthesized lifecycle models from knowledge representation and Information Science Fields. The DELOS lifecycle model is applied (and extended) with regard to two different aspects of TS, a) the vocabulary as an entity in itself and b) elements of a vocabulary used as part of an information system (eg providing a search term). In the latter case, a search system might be 'terminology-aware' in its use of TS, or it might simply treat terminology elements as a source of uncontrolled terms, for its purposes. These two aspects are combined together informally in the revised framework - it should be emphasized that other configurations and selections of the elements are possible. The purpose is to provide a heuristic, unifying framework for considering the range of TS applications. In Section 5.4, some projects are roughly located within this broad framework.

The TS Lifecycle framework is given in Figure 1. Creation here refers to the production of a vocabulary, while Acquisition refers to the stage when the vocabulary is integrated with a collection or a registry of some kind. Identification (considered under Cataloguing) provides a unique key for a vocabulary or a vocabulary element (see Section 6.3). Integration is discussed in Section 4.4. Access, Search and Discovery has been treated in more detail than the DELOS version due to the focus on TS. Of course, the other elements are also relevant to this review, with Acquisition, Maintenance and Archiving being rather less central. Note that the lifecycle may involve creators/authors, publishers, information systems managers, service providers and end-users of different kinds. More generally, this lifecycle model connects or overlaps with wider models of information seeking behaviour (see Section 4.1) and the scholarly lifecycle (Lyon 2003). Note that in practice many user activities involving TS are an iterative process.

Creation and modification of vocabularies
Creating/sustaining vocabularies

Publication of vocabularies
Licensing

Acquisition of vocabularies
Selection, storage

Cataloguing (metadata, identification/naming, registration)
Indexing/classification/annotation

Intellectual, semi-automatic, automatic, disambiguation

Integration (syntactic and semantic interoperability issues)
Mapping, merging

Mediation
User interfaces (TS surfaced in interface)
Faceted/Spatial/Other access afforded
Browsing and visualization of vocabularies
Personalization (of interface or vocabulary)

Access, search and discovery
Discovery of services, vocabularies (and concepts), databases/collections,
Search and retrieval
Querying
TS-aware[2] query
Query expansion – synonyms, semantic
Cross-searching, cross-browsing across distributed collections

Use (as part of a broader service)
Search+Analysis applications,
Information extraction, mining

Maintenance of vocabularies
Evolution, versioning
Application specific, collaborative work

Archiving and preservation of vocabularies

Figure 1. TS Lifecycle framework

4.3 Types of Terminology Web Services

This section discusses terminology services, in relation to the JISC Information Environment and e-Framework. We first discuss Terminology Services as abstract services and then go on to discuss hierarchical layers of services and Terminology Web Services, specifically.

In the context of a service oriented view of the JISC Information Environment, where information services, at least in part, may be offered through web services for use by software applications, it is necessary to extend the scope of terminology related services

[2] Different kinds of vocabularies can serve as the (one) source of query terms. By *TS-aware* query system, we mean that some use is made of the vocabulary by the search software, as opposed to search systems that do not take any of the vocabulary context into account. Examples include fielded search based on type of concept, query expansion based on KOS semantic relationships, (semi) automatic disambiguation in resolving an initial free text term to a controlled vocabulary concept etc.

beyond the few, initial examples documented to date, e.g. by Andy Powell (Powell 2005b, Nov). Terminology services are also mentioned in other service oriented architectures and e-framework initiatives such as those adopted by the DLF Abstract Services Taskforce in the US (DLF) and the e-Framework for Education and Research in the UK (e-Framework), including its component, the E-Learning Framework (ELF). Again, TS are not delineated in any detail and one of the aims of this section of the review is to outline some of the different types of TS, with a view to stimulating more detailed design work in this area.

Recommendations:
Develop more precise definitions of TS, as part of the JISC IE and eFramework
Define search process workflow of TS within JISC IE eFramework

At the service components level, as shown in the JISC IE architecture diagram (figure 1 in Powell 2005b, Nov.), a group of services forms the Terminology Services component. Four are subsequently listed as abstract services: *Vocabulary search interface, Vocabulary harvest interface, Vocabulary deposit interface and Terminology service (*mapping and expanding terms). (News channel and Delete interface are also mentioned without details). Relevant vocabulary standards and protocols should be involved in any bindings of the abstract services. However, it is not clear what precisely is entailed by these services.

4.3.1 Definition of Terminology Web Services

A more comprehensive definition of terminology web services is the following:

> **Terminology Service**s are a group of abstract services, presenting and applying vocabularies, their member concepts, terms and relationships, describing the meaning of terms and facilitating semantic interoperability. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc.

Potentially, abstract services supporting creation, storage and management of terminology might be added, such as *deposit, manage, edit, delete.* They may partly overlap with services for presenting and applying vocabularies, but are not at this stage included in the definition and examples below.

Note that in this context, vocabularies include the different types of controlled vocabulary described in Section 3.1 and, additionally, sets of mapped (or translated) terms and concepts resulting from mapping services. Uncontrolled vocabularies, such as uncontrolled term lists, author provided keywords, tagsets, folksonomies should also be included for terminology service purposes.

There are layers of services at different levels of granularity. At the bottom level, **bindings** are particular instantiations of an abstract service, giving (as appropriate) specific data representations, an API and **Web Service** specification (if that is the

architecture adopted). At higher levels, the abstract services will form part of broader application services or JISC **Service Components**. Abstract services may involve layers of Terminology Services, for example a Search Interface or Harvest Interface might take keywords from a lower level Terminology Service.

Recommendation: Within the context of eFramework develop a hierarchical layered set of protocols for TS and standard bindings to (various) APIs

For each service, various standards and protocols apply. They generally fall into two contexts:
a) standards relating to the description, structuring and functions of vocabulary systems/schemes themselves (eg SKOS Core, VDEX, Zthes, SKOS API, BSI and NISO standards, etc.) – see the various vocabulary related standards described in Section 6.
b) standards related to the TS application context: searching, harvesting, alerting, and other abstract services.

Not all possible bindings and combinations of services are considered here. Some key reference implementations for different types of terminology service would be a useful future development.

As an initial step, the next section provides more detail on selected TS. It builds on and extends previous efforts, which deal with terminology services with broad brush strokes (eg Powell 2005b, Nov). However, it is still far from complete. The distinction between 'business processes' and individual abstract services (Powell 2005a, Feb) has not been followed, since this adaptation to the DLF approach is not widely adopted at this time.

4.3.2 Groups (and layers) of abstract terminology services

Three broad groups of abstract terminology services are described below (the third in less detail). In order to illustrate that hierarchical layers of terminology services are necessary, lower level terminology service options for some relevant cases are detailed (in italics). These would be called as lower level services, as part of the implementation. The SKOS API (see Section 6.4.1) is used to express the low level terminology services. This has a Java Web Service binding but is also expressed as a binding-independent protocol (and thus could have an HTTP implementation, say). In the listing below, they could be considered, more or less, as both a specification of a low level terminology service and one possible binding of it. Another binding of the same low level terminology service would be possible. Some lower level OAI harvesting service examples are also shown.

The first of the three groups concerns abstract services related to entire vocabulary schemes/systems. The distinction between discovering (identifying) a suitable vocabulary and retrieving metadata about it, versus retrieving member concepts and terms of a vocabulary tends to be overlooked. The differentiation between services relating to a complete vocabulary scheme and its metadata versus services relating to member terms (the second group below) is fundamental, as is the differentiation with services related to the application (more or less seamless) of terminologies in other services.

For example, an abstract service described as 'Vocabulary harvest interface' (Powell 2005b, Nov) does not specify whether it concerns only the 'harvesting' of individual terms and other information about the vocabulary, or groups of terms, or perhaps the harvesting of the complete vocabulary. (The last would tend to be an exception, considering current practice and the rights situation regarding vocabularies.)

Another intended contribution of the description below is to illustrate that terminologies may comprise entities other than terms: i.e. concepts and relationships, and that services can serve such entities from either one or several different vocabularies.

1 Services related to the vocabulary (encoding) systems/schemes

11 discover suitable scheme in vocabulary registry
and:
search, browse, harvest, alert, upload/deposit, edit etc.

12 disclose selected or complete information (metadata) about scheme(s)

getSupportedSemanticRelationsByThesaurus(URI thesaurus)

13 statistics (e.g. information about size and usage levels)

2 Services related to member terms/concepts/relationships from one or several vocabulary systems

21 discover/search member
term/concept/relationship/translations/mappings/structures
[(authority) look-up] in one or several vocabulary systems
and:
search, browse, harvest, alert, upload/deposit, edit etc.

getConceptsMatchingKeywordByThesaurus(keyword, URI thesaurus)
getConceptsMatchingRegex(regexp)

22 disclose or harvest terms/concepts/relationships/translations/mappings/structures
(known item)

getConcept(uri)
getConceptByExternalID(externalID, URI thesaurus)
getConceptByPreferredLabel(preferredLabel, URI thesaurus)

23 browse in networks of terminology

24 disclose subsets of the topological environment of terms/concepts/relationships; several terms/concepts and semantic relationships between them; a synset;

translations; mappings; subsets of hierarchies; a concept and all related terms; a classification and all related information; a term and all translations; etc.

getSupportedSemanticRelationsByThesaurus(URI thesaurus)
getAllConceptRelativesByThesaurus(concept, URI thesaurus)
getConceptRelativesByThesaurus(concept, relation, URI thesaurus)
getConceptRelativesByPath(concept, relation, URI thesaurus, int distance)

getTopConcepts(concept, URI thesaurus)
getTopmostConcepts(URI thesaurus)

(OAI) harvest of sets

25 harvest a complete vocabulary/mapping set

(OAI) complete harvesting function

26 upload/deposit member terms/concepts/relationships/translations/mappings

27 edit member terms/concepts/relationships/translations/mappings

28 alert about new or changed member terms/concepts/relationships/translations/mappings

3 Services related to the application of terminology in other services
(a rough illustration)

31 automatic indexing
32 term or keyphrase extraction
33 named entity recognition, data mining
34 automatic translation of term or document
35 query enhancement, query expansion
36 automatic classification
37 automatic mapping
38 semantic reasoning
...

Vocabulary searching and browsing functions, integrated into an information service, are not fundamentally different from external terminologies used for these purposes (i.e. the abstract terminology services in group 2 above). Technically, they could be invoked as (web) services. Even the suggestion of terms from a controlled vocabulary can be accomplished by sending a suitable request to a web service such as 21, 22 or 24.

It is doubtful whether vocabularies completely integrated within a data or document collection, to the extent there is no separate representation or access, could be the basis

for useful web services. However, the performance of external web service calls is a critical issue for integrated, responsive search and browse interfaces. It remains to be seen whether and which kind of web service architectures will deliver this kind of response. In some circles, there is interest in light(er)weight approaches to web services (eg SRU and REST protocols), due to concern about the overhead imposed by SOAP-based approaches for interactive Web applications, when response time is an issue. This is also taken up in Section 6.4.1 in the discussion of protocols, specialised for user-facing terminology services. The next section gives an illustrative example of layered terminology services from group 2 above, in context of terminology-assisted search.

Recommendation: Develop open source, reference terminology web service implementations

4.3.3 Illustration of TS assisted search process

This section gives a breakdown of some of the detailed steps in the terminology-aware search process, as a further illustration of the layers of services that can be involved in searching and the iterative nature of that process. Figure 2 attempts to combine search and browsing operations. It assumes that vocabularies are in a Registry and associated with specific Collections. An initial user search statement needs to be expressed as controlled terminology, either via browsing, or by resolving the initial terms into controlled terms (disambiguating if necessary). In some cases, browsing can trigger a query, otherwise a query is formulated with concepts from the vocabulary. Results can be inspected for query reformulation purposes and different forms of query expansion can be applied. Saved queries or results can form the basis for starting the process over again. The diagram is a simplified version of the reference thesaurus search model, discussed by Blocks *et al.* (2006). The ultimate aim would be to map the lower level terminology services, outlined in the previous section, into the diagram's search process 'work flow'.

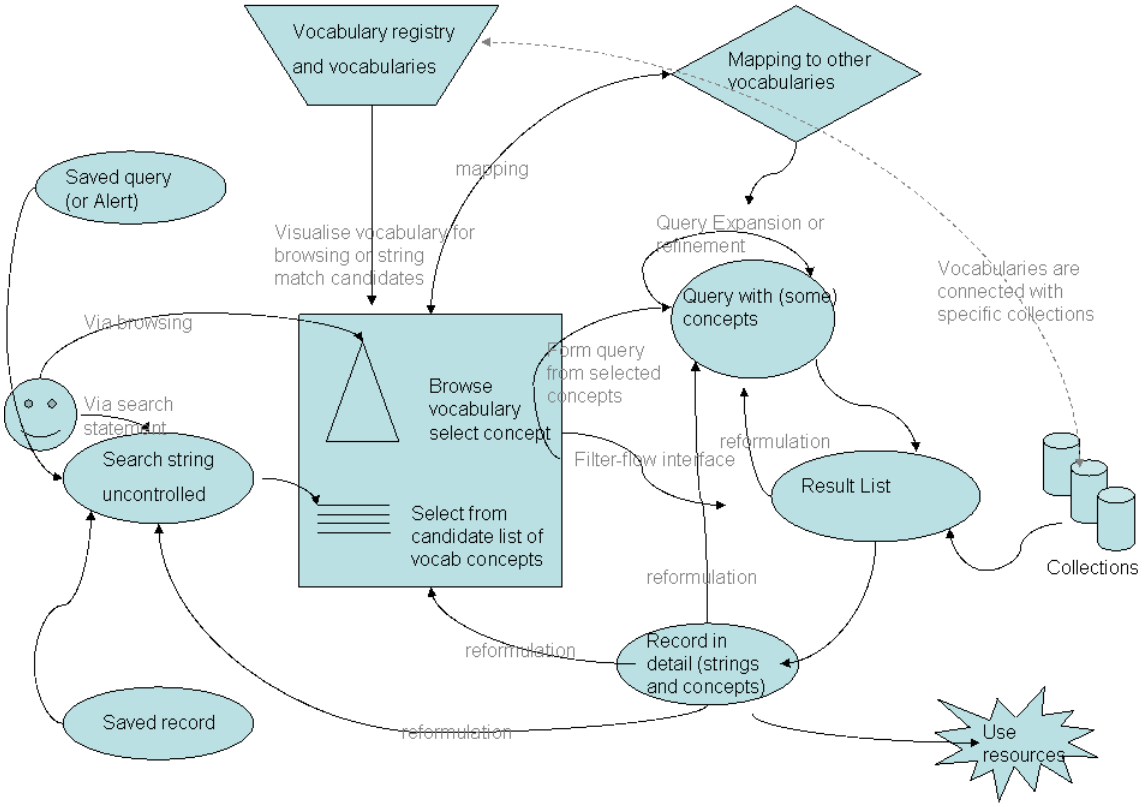

Figure 2. Illustrative diagram of controlled terminology search process. Simplification of diagram in (Blocks *et al.* 2006)

4.3.4 Terminology Web Services review

This section contains a brief review of work to date on terminology services with web service bindings.

The OCLC Terminology Services project is described in Section 5.3.6. Salient aspects are briefly reviewed here. Multiple vocabularies are made available via a range of web services (based upon SRW/SRU with the Zthes profile, the MS Research Pane and REST and SOAP protocols). Encoding formats are HTML, MARC-XML and MARC21 Authorities or Classification formats. Other OCLC web services are the Name Authority Service as an add-on to DSpace and the LAF (LC Name Authority File) web service which was developed in the context of a collaboration with the ePrints UK project (cf. section 3.3) (LC Name). Vizine-Goetz (2003) discusses OCLC mapping services, while Vizine-Goetz *et al.* (2006) describe their SOA architecture and use of Microsoft's Research Pane.

HILT3 has plans to develop web services based upon SRW and SKOS Core (see Section 5.1.4). See also section 3.2.4 on the Becta Vocabulary Studio, which provides a web service interface to its Vocabulary Bank.

Various web services activities form part of a wider ECOinformatics Initiative (http://ecoinfo.eionet.europa.eu/), which has seen Environmental Thesaurus and Terminology Workshops. The CSA/NBII Biocomplexity Thesaurus Web Services project have adapted the SKOS API to search and browse the Biocomplexity Thesaurus. Not all functions have been implemented but an efficient keyword search element has been added (CSA/NBII). A multilingual demonstrator is also available. The GEMET multilingual environmental thesaurus was developed by EKOLab using the T-REKS (Thesaurus-based Reference Environmental Knowledge System) model. EIONET (European Environment Information and Observation Network) have developed a web service API for accessing an extended version of GEMET, using an SKOS representation (GEMET). Tudhope and Binding (2006) discuss initial experiences with a web service browser, using a restricted set of the SKOS API functions.

The California Environmental Digital Library Network (CalEDLN) uses a web service API for interaction between SKOS encoded vocabularies and metadata editing and search interfaces. The API supports access and browsing of distributed vocabularies, term matching and thesaurus functionality for thematic keywords, California place names, resource types, and person/organization names. The software solutions are open source (CalEDLN). This is a further development of the CERES Protocol work mentioned in Section 6.5.1.

DLESE has developed a NSDL Strand Map Service as a Networked Knowledge Organization and visualization system for K-12 Education (the REST-based web service protocol generates the visualizations) taking the form of an enriched end user thesaurus. It invites both teachers and learners to make connections between important ideas (Sumner 2005).

MelvilSoap is a web service from Die Deutsche Bibliothek. offering DDC 22 in German. It allows users to query Melvil, the German WebDewey, via a SOAP interface. It is intended to be used in classification work and to support searching German databases carrying DDC classification. The service requires subscription (Melvil).

MeSHine (EUTROPA.de) is a web service using MeSH in German and other languages to search the Internet (Google web-APIs) and Pubmed/Medline (using the Entrez-SOAP utilities of NLM – see also Section 5.2.5); SOAP/REST access to other vocabularies can be developed. The vocabularies are structured in XML messages format (MeSHine).

Zisman *et al*. (2002) discuss experiences from applying Web service wrappers in an 'information bus' approach to the development of a prototype system that integrated various UN FAO data sources with disparate organisation and structure.

There is some current work looking to integrate web services more fully with Semantic Web efforts. There is a W3C Semantic Web Services Interest Group, with enhanced semantic description and choreography of web services. The ongoing Web Services and Semantics Project (IST-FP6-004308) is an EC specific support action in this general area.

Recommendations:
Collaborate with international efforts in terminology web services
Develop a range of TS-based search and browsing tools

4.4 Mapping

Mapping is a key requirement for semantic interoperability in heterogeneous environments. Although schemas, frameworks and tools can help, detailed mapping work at the concept level is necessary, requiring a combination of intellectual work and automated assistance. Significant effort is required for useful results.

Although some major integrated vocabularies exist (for example, UMLS, GEMET), it is often the case that construction of purpose built integrated vocabularies is not practical. Accordingly, mapping between vocabularies is important for facilitating access to information resources in different contexts, different purposes and for different user communities.

It is sometimes possible, however, to map to an appropriate switching vocabulary. There has been experience with projects, such as HILT and Renardus, mapping to DDC as a central spine (eee Section 5). The Becta Vocabulary Studio, dynamically maps terms in its Vocabulary Bank to a central spine of concepts (see section 3.2.4). There is also ongoing research into employing top level core ontologies as integrative frameworks between different domain vocabularies and heterogeneous datasets within broad domains (see eg Doerr *et al.* 2003 and the CIDOC CRM, developed for the museum community and being extended to libraries and archives). It is also possible to make us of linguistic resources, such as lexical databases (eg WordNet) and linguistic ontologies (eg Navaretta *et al.* 2006), to assist mapping efforts.

The DELOS Report D5.3.1: Semantic Interoperability in Digital Library Systems (Patel *et al.* 2005, section 6) discusses these issues in more detail. It compares information science and ontology-based mapping methodological approaches, concluding they are intrinsically fairly similar. The HILT project reports also review different mapping approaches.

Zeng and Chan (2004) provide an extensive recent review of mapping work. They identify several methodological options, prominent among these being:

a) **Derivation/Modeling** of a specialized or simpler vocabulary from an existing complex vocabulary. For example, facet analysis can play a key role in facilitating semantic interoperability by deconstructing and systematising complex, pre-coordinated Subject Headings that might otherwise prove intractable for mapping purposes. The OCLC FAST project (FAST) has converted LCSH headings via a simplified syntax into a faceted representation.
b) **Translation/Adaptation** from a vocabulary in a different language.
c) **Satellite and Leaf Node Linking** of a specialised thesaurus to a large, general thesaurus. This is a cost effective method for augmenting a widely used general vocabulary with more specific local concepts and terms. In time, the additions

may be adopted by the general thesaurus editors but in the meantime they should always be identified as local.

d) **Direct Mapping** between concepts in different controlled vocabularies, usually with an intellectual review.
e) **Co-occurrence Mapping** between two vocabularies based on their mutual occurrences within the indexing of items within a collection. Co-occurrence mappings are considered looser than direct mapping made by experts.
f) **Switching** language used as an intermediary. It can be a new system (e.g. UMLS Metathesaurus) or an existing system. This is one of the most frequently used approaches, see for example use of the DDC in projects HILT and Renardus.

Of course, there are also variants and combinations of these approaches. In practice, the success (and cost) of a vocabulary mapping operation will tend to depend on the congruence of the vocabularies to be mapped. Relevant factors include the degree of overlap, degree of pre/post-coordination, similarity in structure and level of specificity, the target application and context of use (for more details, see the discussion in Patel e*t al.* 2005, Section 6.2.1).

Some significant projects have worked in this area – see Section 5 for outlines of HILT, OCLC, Renardus projects. The UN FAO are investing resources into mapping work with the Agrovoc thesaurus (Liang and Fini 2006). The initial HILT project concluded that one high level vocabulary was not feasible for JISC purposes and has piloted terminology services at the collection level for UK higher educational communities, via mapping to a DDC spine. Similarly based on DDC, the Renardus project created a common 'switching' structure to support a cross-browsing service (Koch *et. al* 2003). OCLC (providers of the DDC) have developed several mappings between major vocabularies (both intellectual and statistical), now available as terminology web services (OCLC Terminology Services, Vizine-Goetz *et al.* 2003). The OAI protocol is used to provide access to a vocabulary with mappings, via a browser to human users and through the OAI-PMH web service mechanisms to machines. Both direct mappings and co-occurrence mappings are provided, depending on the situation.

Part 4 of the draft BSI Standard on Structured Vocabularies is concerned with interoperability and mapping between vocabularies and gives some useful examples, both mono and multilingual. It also has a discussion of the impact on retrieval of different options. This is an important consideration, particularly when no exact equivalent concept exists, and it is necessary to map to a broader or narrower concept, a partially overlapping concept, or to a (Boolean) combination of concepts. It distinguishes mapping for index terms, search terms, pre-coordinated strings, one to many, many to one mappings, etc. Different types of mapping relationships and types of inexact equivalences (partial mappings) are discussed. Set-based approaches to mapping are outlined by Renardus (Koch *et al.* 2003), with regard to classification schemes. (Note that there may be differences in mapping approaches for different types of KOS, eg classifications versus thesauri). Doerr (2001) proposes an extended set of mapping relationships and discusses mapping issues generally. This was an influence on the draft SKOS-Mapping Schema, which describes RDF thesaurus vocabulary extension for defining inter-

thesaurus mappings and equivalence relationships, although it has yet to see serious application to evaluate its proposals.

Recommendations:
Investigate and compare different mapping approaches and granularities in pilot projects
Develop a range of TS-based tools to assist in creating mappings
Investigate the potential for standard mapping relationships and a mapping protocol
Collaborate with international efforts in mapping services

4.5 Automatic classification and indexing

Automatic classification and indexing (see Section 3.2.1.1 on the distinction) tools are important for addressing the potential resource overheads in applying TS to indexed collections and repositories. Some tools are emerging that should be investigated for JISC purposes. Many argue that a combination of intellectual and automatic methods is currently an optimal approach (eg Hagedorn 2001). Human input can be used to design vocabularies used by subsequent automatic stages and can also intellectually review automatic results.

In a recent review of automatic subject classification methods, Golub (2006a) distinguishes three discipline-based approaches: text categorization using AI machine-learning techniques; document clustering using (information retrieval) statistical techniques; document classification using controlled vocabularies. Analysis of citation patterns reveals that the three approaches have tended not to overlap. However this may now be changing. Medelyan and Witten (2006), from the University of Waikato, report on a combination of thesaurus-based indexing with naïve Bayes machine learning methods for domain-specific keyphrase extraction that achieves results close to the inter-indexer consistency found in professional human indexing. Their new Kea++ algorithm is available under an open source license.

In a review for HILT, Russell and Day (2001) briefly review some commercial automatic classification tools: Autonomy, Interwoven, Semio, Wordmap. Other commercial products include Collexis with its automatic "fingerprinting" and OCLC's Connexion interactive cataloguing software. The JCDL 2006 workshop on metadata tools for digital resource repositories provides a list of exhibitors, some of whom offer indexing tools. Lancaster (2003) is a standard text on vocabulary based indexing and classification generally.

Various research projects have explored vocabulary-based subject classification and some automatic tools are freely available. Larson (1992) conducted early experiments using the Library of Congress Classification. OCLC's longstanding automatic classification project has also investigated automatic web page classification using the DDC and Library of Congress Classification. Their Scorpion project applied a text web page as query to the DDC knowledge base and the resulting tools have been used as

classification support in CORC and OCLC's Connexion cataloguing software. A Scorpion demo and software is available under a research license (OCLC Automatic Classification).

The iVia/INFOMINE project at UC Riverside have experimented using LCSH with machine learning based on a large training set. Paynter (2005) discusses corresponding evaluation methods and tools. The iVia and DataFountains tools for focused crawling and automatic classification are available under an open source license.

Golub (2006b) investigates the problems faced in applying KOS to text-based subject classification of Web pages. A selection of mis-classified Web pages is analysed in great detail to uncover why the automatic methods assigned inappropriate classes and illustrative examples are discussed. The underlying method combines a classification scheme with a corresponding thesaurus to give a rich set of resources for the algorithm. The techniques are based on the automatic classification approach developed by the DESIRE project for a subject gateway in the Engineering domain (Koch and Ardö 2000). They are now being applied by the University of Lund in the EC ALVIS project. Various demonstrators and tools are available from Lund's KnowLib (Knowledge Discovery and Digital Library Research) Group, applying terminologies to classification of harvested fulltext web documents. The ALVIS project offers open source tools with automatic topic classification, including DESIRE's COMBINE (the Combine Harvesting Robot, "an open system for crawling [harvesting and threshing (indexing)] Internet resources"), used by the Swedish web archive.

Recommendation:
Investigate semi-automatic solutions to indexing and classification in pilot projects
Investigate currently available tools for automatic indexing and classification

4.6 Text mining and information extraction

Vocabularies of all types play a prominent role in text and data mining as well as in information extraction tools and services. This is discussed in Section 3.3.3, with regard to name authorities. In some cases automatic classification methods overlap with what is regarded as text mining (see previous section). In this section text mining is considered in a wider context.

Text mining covers a range of approaches to extraction of textual information, ranging from what might be considered 'algorithmically enhanced indexing' through to hypothesis testing. Text mining has particular application for assisting scientists to automatically extract information from the large bodies of text that are now being produced in scientific disciplines (Ananiadou, 2005). There is potential for using existing terminologies and ontologies as auxiliary tools to support text mining. In a reciprocal way it is possible to use text mining as a means to automatically update and expand existing ontologies (through techniques such as term clustering).

Recent advances in language engineering have made available a range of tools that can assist information extraction from free text documents. These include lexical databases (eg WordNet), part-of-speech taggers, parsers and other tools. Realising the potential for TM requires increased availability of large corpora with context data for extracting data using statistical methods. Lynch explores how open access is a probable prerequisite for large scale computational approaches to the scholarly literature (Lynch, 2006).

The JISC funded (in part) National Centre for Text Mining (NaCTeM) is operated by Manchester and Liverpool Universities and has various international collaborators. The initial focus is on bioscience and biomedical texts. A range of resources and links to lexical databases, ontologies, tools, tutorials and open source software is available and services are anticipated via a Web-based portal. Some tools already have been made available including ATRACT and CAFETIERE.

Text mining makes explicit use of ontologies, both to annotate existing terms within texts and to refine the ontology. In order to exploit existing terminologies, it would be worthwhile investigating whether existing thesauri and other KOS would be sufficiently rich to be regarded as 'ontologies' for this purpose, and thus both contribute to text mining processes and, in turn, be enhanced through text mining. Joint working between inter-disciplinary teams has also been evident within the biomedical community where there has been some effort to bring together those constructing ontologies with those researching text processing, and it would be useful to widen such collaboration to other disciplines.

Recommendations:
Investigate relationship between KOS and text mining:

- **Demonstrate how KOS can support text mining**
- **Demonstrate how text mining can be used to update and enhance KOS**

4.7 General sources for work in TS

Beyond JISC sponsored publications, such as Ariadne, the ECDL, JCDL and DCMI conferences are good general sources. The series of NKOS workshops on Networked Knowledge Organisation Systems and Services are another good source for current work in terminology services. NKOS-related special issues have appeared in the online Journal of Digital Information (Hill and Koch 2001, Tudhope and Koch 2004) and the New Review of Hypermedia and Multimedia (Tudhope and Nielsen 2006). See particular sections of this review (eg Section 3.2.4 on eLearning), for some general sources in these areas.

5 Review of current terminology service activity

See also the reports by the JISC Pedagogical Vocabularies Project for eLearning projects and the JISC Shared Infrastructure Services Review.

5.1 JISC related activity

5.1.1 Archaeology Data Service (ADS)

The ADS provide various operational services in the archaeology domain.

5.1.1.1 ADS ArchSearch

http://ads.ahds.ac.uk/catalogue/
The Archaeology Data Service (ADS) offers a map-based search to archaeological resources.

5.1.1.2 HEIRPORT

http://ads.ahds.ac.uk/heirport/
The Historic Environment Information Resources Portal (HEIRPORT) integrates historical data from different public bodies, again via a map-based interface.

Digital Archaeology has seen increasing use of the Web to disseminate data and reports. For example, the HEIRNET portal offers Z39.50 search across a wide range of heritage information via an attractive map-based interface, while the ADS online catalogue, ArchSearch provides similar access to archaeological investigations. These are operational systems. To date, only fielded Boolean search is possible and terminological tools have not been investigated, although further development of spatial and filter flow interfaces in the ongoing Common Information Environment project is outlined below.

5.1.1.3 CIE demonstrators

http://www.common-info.org.uk/index.htm
http://www.ariadne.ac.uk/issue39/miller/
The Common Information Environment (CIE) is a cross sector collaboration, aiming to promote a common infrastructure for accessing information and opening up access to the hidden web of databases and collections (Miller 2004). JISC funded two technical demonstrators which each integrated resources of different types from several collections. Adiuri Systems developed a Health Demonstrator using their faceted, filter-flow interface with medical vocabularies such as SNOMED and MeSH and various collections, including the RDN hub, BIOME. The Archaeology Data Service (ADS) led a collaboration to develop a Place-based demonstrator, building on experience with HEIRPORT (and contributing to further enhancements of HEIRPORT), RCAHMS and also Edina's geoXwalk (http://www.britarch.ac.uk/HEIRNET/cie_demonstrator.html). Map-based search and presentation of results was supported.

Subsequently ADS and Adiuri Systems were contracted to deliver an enhanced Place Demonstrator with a faceted interface, sophisticated presentation of results and portlet capability (http://www.common-info.org.uk/enhanceddemonstrator.htm).

5.1.2 Co-ODE: Collaborative Open Ontology Development Environment

http://www.co-ode.org/

Collaborative Open Ontology Development Environment (Co-ODE) is a Semantic Grid project funded by JISC at Manchester University to develop freely available and easy-to-use Ontology management and OWL tools, as plug-ins to Stanford University's Protégé. The latest download is a debugging tool. A set of tutorials are also being developed. The project builds on earlier experience with the lightweight ontology editor OilEd. HyOntUse (User Oriented Hybrid Ontology Development Environments) is a more theoretical sister (EPSRC eScience) Manchester project, concerned with issues such as debugging ontologies.

5.1.3 geoXwalk Gazetteer Service

http://hds.essex.ac.uk/geo-X-walk/

Geo-spatial Gazetteer Service is a collaboration between Edina (Data Library, University of Edinburgh) and the UK Data Archive (University of Essex) to provide a JISC shared service. It offers feature (concept) searching, together with geographic searching and spatial operators and syntactic geographic term mapping. Results are presented in context of a map-based spatial visualisation, with flexible footprint options for result items. The aim is that (legacy) implicitly geographically referenced resources may be made explicitly geographically searchable via a GeoParser. Thus the service automatically indexes resources via geographically-specific information extraction.

This third project builds on two previous JISC gazetteer projects and aims to take the previous Demonstrator on to a full shared service. The geoXwalk gazetteer builds on and adapts the ADL Gazetteer Content Standard and ADL Feature Type Thesaurus. Phase 2 employed the HTTP ADL Query Protocol.

5.1.4 High Level Thesaurus (HILT)

http://hilt.cdlr.strath.ac.uk/index2.html

HILT is concerned with facilitating subject-based access across the broad provision of JISC collections and automatic discovery of relevant collections. Two previous and one ongoing HILT project phases have investigated pilot terminology services, in collaboration with OCLC Research and the company, Wordmap. Starting off as a pilot project to investigate the feasibility of a High Level Thesaurus for HILT DDC was chosen as a central spine for mapping between vocabularies (particularly DDC, LCSH, UNESCO, MeSH, AAT), making use of OCLC Research's available mapping services. The pilot HILT2 demonstrator offered a cross searching facility via collection level DDC descriptions. A user term indicating the subject of interest is mapped to the DDC terms (intellectually disambiguated if necessary) and this is used to suggest a set of relevant JISC collections. The current system allows only single user terms. Successive truncations of the DDC number (to more general concepts) are applied if there is a failure to match any collection. The original DDC number is then automatically mapped to the vocabulary used by the collection (if one of the mappings covered) and where possible an automatically search is conducted. HILT3 is currently developing a M2M demonstrator based on webservices, the SRW protocol, SKOS Core and SKOS-type concept URIs. It is

planned that end-users will not access HILT directly but via web-based user services, such as GoGeo! and BIOME.

5.1.5 Learning and Teaching Portal (Portals Programme)

http://www.connect.ac.uk/

The Connect Learning and Teaching Portal was jointly funded by JISC and LTSN (Learning and Teaching Support Network now part of the Higher Education Academy). It aims to facilitate resource discovery of information on organisations, funding opportunities, projects and sector resources on learning and teaching research. Systems Simulations Ltd (and their Index+ retrieval system) was contracted to deliver the portal. Information items have been indexed with six educational vocabularies: Educational Level, JACS, LearnDirect, RDN/LTSN Resource Type, Pedagogy, Policy Themes and Region. Facilities include fielded text search and search/browsing via the vocabularies. The portal has been designed as a set of discrete resources that can be embedded in other portals or websites.

5.1.6 Mersey Libraries, Archives Hub and Cheshire

http://www.merseylibraries.org/about.html
http://www.archiveshub.ac.uk/index.html
http://cheshire.berkeley.edu/

MerseyLibraries.org provides access to local distributed collections via the (University of Berkeley) Cheshire Information Retrieval system and Z39.50. The Archives Hub provides access to descriptions of archives held by Universities and Colleges. It ius funded by JISC and hosted at MIMAS, with systems development by University of Liverpool.

Cheshire II is an operational online catalog and full-text retrieval system with ranked results via probabilistic Information Retrieval (IR) techniques. It supports fielded search, relevance feedback, hypertext linking. Cheshire3 has support for SRW/U, CQL, OAI and XML namespaces and has a documented API. Cheshire's Entry Vocabulary Indexes support the mapping from free text to controlled vocabulary terms, applying various techniques to yield a ranked list of terms. In addition there is support for Z39.50-based thesaurus and gazetteer searching.

5.1.7 Resource Discovery Network (RDN)

http://www.rdn.ac.uk/

The longstanding and influential JISC RDN is a set of subject gateways to selected Internet resources for learning, teaching and research. As of summer 2006, it will be reorganised into the Intute service, with resources held centrally. Z39.50, SRW and CGI interfaces are provided. The RDN-Include mechanism allows the top-level browse structure and basic search facility to be embedded into a Web-site.

Different hubs have different subject coverage and currently have different local retrieval systems. Depending on the hub, resources can be indexed by a domain thesaurus, covering wide areas within a broad subject, such as AAT, CAB, MeSH, and/or classified under broad headings or by DDC, LC Schedules. Some hubs are indexed according to the LTSN / RDN (RLLOMAP) LOM application profile and associated vocabularies.

With regards to TS, typically, hubs provide browsing access via a vocabulary (sometimes presented alphabetically) and free text search. The search functionality is not terminology aware but search results can include controlled terms, which can yield further resources.

5.2 Other UK activity

5.2.1 COHSE Conceptual Open Hypermedia Project

http://cohse.semanticweb.org/

COHSE was a joint University of Manchester – University of Southampton project, funded by the EPSRC. The aim was to use ontologies to help improve the range of hypertext linking possibilities, building upon experience with Southampton's Microcosm dynamic linking techniques. Web pages were automatically annotated with concepts, dynamically inserted based on words in the text. This formed the basis for hypertext links to related content and subsequent interactive navigation of the web pages. An evaluation was conducted with a Java programming tutorial at Sun, using a purpose-built ontology. As part of the project, OilEd, an open-source downloadable, lightweight ontology editor for DAML+OIL was developed.

5.2.2 FACET

http://www.comp.glam.ac.uk/~FACET/facetproject.html

FACET was an EPSRC funded project by University of Glamorgan in collaboration with the National Museum of Science and Industry (NMSI). The aim was to investigate the integration of a thesaurus into the search system and the potential of faceted thesaurus-based query expansion techniques. An extract of the NMSI collections database formed the testbed. A Web demonstrator is available, illustrating dynamic control of thesaurus-based query expansion parameters and producing ranked results for multi-concept queries. The ranking is based on conceptual distance in the thesaurus, allowing for difference in choice of concepts by indexer and searcher.

5.2.3 FATKS

http://www.ucl.ac.uk/fatks/
http://www.ucl.ac.uk/fatks/database.htm

FATKS was an AHRB funded project by University College London to investigate the method of facet analysis in the humanities. A faceted classification scheme was developed, building on Bliss Bibliographic Classification 2, Universal Decimal Classification and the Broad System of Ordering. This afforded free combination of concepts in indexing. It was applied to a Web demonstrator in the areas of religion and

visual arts, which illustrated searching and browsing of the classification. Facet analysis techniques investigated include facet citation order, synthesis rules, notation/facet indicators, common auxiliaries.

5.2.4 FISH Interoperability Toolkit

http://www.heritage-standards.org/
http://www.fish-forum.info

The FISH Interoperability Toolkit has been funded by English Heritage and the National Trust. FISH (Forum on Information Standards in Heritage) have developed this XML Schema as a common format for 'the storage, processing and exchange of historic environment information'. It supports the FISH MIDAS Standard and includes a set of controlled vocabulary identifiers to identify namespaces for thesauri or terminology lists, as well as a validation tool. The Web Services Historic Environment Exchange Protocol (HEEP) offers a standard route for programmatic access to data conformant with the schema.

5.2.5 NHM Nature Navigator and other Scientific Taxonomic Projects

http://www.nhm.ac.uk/nature-online/biodiversity/nature-navigator/

Various projects at the Natural History Museum (NHM) over the years have made use of scientific taxonomies to organise and present online information. Currently Nature Navigator, funded by funded by the New Opportunities Fund Digitise Programme, "provides a single access point to information on more than 8,000 of the best-known species that occur in Britain. It will guide you through the mass of names of organisms, showing you the preferred scientific and common names, related organisms and where they fit into the classification of the natural world." The Navigator allows browsing access to the taxonomy, expanding and collapsing branches. It juxtaposes scientific and common names, allowing lookup via either method with interactive disambiguation. The taxonomy is integrated with display of the relevant fact sheet (and pictures) on the organism. The taxonomy is based on ITIS (Integrated Taxonomic Information System - http://www.itis.usda.gov/).

The related Species Dictionary Project (http://nbn.nhm.ac.uk/nhm/), in collaboration with the National Biodiversity Network, is developing an exhaustive, standard reference for names of UK organisms from a wide range of datasets. Again, it is possible to search by common or scientific names. Other major scientific taxonomic online database projects include Species 2000 (http://www.sp2000.org/) and the Catalog of Life, which is a collaborative project between Species 2000, IT IS and the Global Biodiversity Information Facility (GBIF). The Catalog of Life (http://annual.sp2000.org/search.php) is intended to provide a validated index to known species in order to monitor biodiversity worldwide. It is searchable and browsable on the Web. The NCBI *Entrez* Life Sciences Search Engine (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) provides a search facility across a range of life science databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy. It incorporates a

Taxonomy Browser. The Tree of Life (funders include University of Arizona and the NSF) project (http://tolweb.org/tree/phylogeny.html) is a worldwide collaboration that aims to provide searchable and browsable information on the evolutionary tree for organisms worldwide and corresponding taxonomic data. Treehouses provide information for "k-16 learners, teachers and the young at heart".

5.2.6 OpenGALEN

http://www.opengalen.org/

OpenGalen is a Manchester University project on medical terminology systems. Based on the high level Open GALEN Common Reference Model, it applies a logic-based approach to medical coding and classification systems. This is delivered via an ontology developed based on a formal knowledge representation language, GRAIL. This predates current Semantic Web description logics such as OWL but is closely related and has specific features for medical modeling. Medical terminology codes have traditionally been enumerated – Galen applies a faceted, synthetic approach to composition of multi-concept descriptors. Given a definition of medical concepts, the system organizes them into hierarchies. Galen has been applied in EC funded projects to classification of basic elements in surgical procedures, with multilingual, natural language descriptions. A GALEN Terminology Server is available and also the OpenKnoME client knowledge management application is available on an open source basis for Microsoft Windows platforms, along with a time limited server.

5.2.7 SKOS (Simple Knowledge Organisation System)

http://www.w3.org/2004/02/skos/

The SKOS work on standards for thesauri and related KOS is an outcome of the EC FP5 SWAD-Europe project, involving CCLRC, ILRT and others. The aim was to facilitate the migration of KOS to the Semantic Web, building on previous work in LIMBER and other projects. The main outcome is the SKOS-CORE Vocabulary and representation in RDF, currently under development within the W3C Semantic Web Best Practices and Deployment Working Group, as a W3C Working Draft. A Guide is available. There are also less stable SKOS Extensions for specializations of SKOS Core. Other ongoing projects are SKOS Mapping, which is looking at using RDF to express mappings between concept schemes and the SKOS API, a Web Service API (and Java implementation) for programmatic access to thesauri. See Section 6.4.1 for further discussion.

5.2.8 STAR (Semantic Technologies for Archaeological Resources)

STAR is a project by University of Glamorgan, in collaboration with English Heritage, recently funded by the AHRC. It seeks to combine query expansion techniques from FACET with the CIDOC Conceptual Reference Model (CRM) as an integrative framework. The aim is to investigate techniques for searching across archaeological databases and linking them to grey literature reports using natural language tools.

5.3 International activity

5.3.1 Alexandria Digital Library

http://www.alexandria.ucsb.edu/
http://www.alexandria.ucsb.edu/gazetteer/
http://www.alexandria.ucsb.edu/thesaurus/
http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm

The Alexandria Digital Library (ADL) was an NSF DL funded Project at University of California, Santa Barbara. It included various components, including ADL middleware and the ADEPT Virtual Learning Environment. The ADL comprises a distributed set of georeferenced collections, with a map-based interface for both input and result display. The middleware includes the influential geographic place ADL Gazetteer Content Standard and the lightweight HTTP ADL Gazetteer Protocol for programmatic access, which have been adapted by the Edina geoXwalk projects. Open source (Java) gazetteer protocol client and matching server are available. Footprints can be defined via the Open GIS Consortium's Geography Markup Language (GML), or other supported geometry languages. There is also a lightweight Java HTTP ADL Thesaurus Protocol (see Section 6.5.1). The ADL Feature Type Thesaurus for types of geographic features has been adapted by Edina. The ADL Gazetteer (5.9 million geographic names) is also available.

The Adept VLE investigated the learning of concepts as an integral element of learning scientific material. In a multi-screen classroom projection, concepts were presented as part of a concept network. ADL technology underpins an ongoing (Library of Congress Digital Preservation Initiative) geospatial library network project.

5.3.2 E-Biosci : EC platform e-publishing and info integration in Life

http://www.e-biosci.org/

E-Bisosci is an EC collaborative project, led by the European Molecular Biology Organisation, which aims to provide access to information in the Life Sciences, linking genomic data to the life sciences research literature, including Medline, Ingenta, BioMed central, PubMed Central, Nature Publications. Collexis technology is used for indexing and thesaurus-based search with multi-concept ‘fingerprints’, where relative importance can be adjusted. Queries can initially be keywords or extracts from abstracts. Matching concepts are shown in results (of scientific articles) and can be used to refine the search. Links are shown from abstracts to genes and genetic databases.

5.3.3 Renardus

http://www.renardus.org/

Renardus was a European project, involving Netlab (University of Lund), UKOLN, ILRT and other partners. A multilingual interface provides a cross-browsing service, with integrated search and browsing access simultaneously across distributed subject gateway services from participating European providers. This is available for use on the Renardus

website. The local gateways classification systems (and browsing structures) are mapped to a common central spine, the Dewey Decimal Classification (DDC). The upper levels of the DDC system forms the basis of the central browsing display, with over 2200 links to subject gateway collections and their browsers. A cross-search facility is also provided. See also the automatic classification tools in the DESIRE Project (see section 4.5 for Koch and Ardö, 2000)

5.3.4 Simile Piggy Bank

http://simile.mit.edu/piggy-bank/
The MIT and W3C SIMILE (Semantic Interoperability of Metadata and Information in unLike Environments) project is concerned with general interoperability in Semantic Web environments amongst data, metadata and vocabularies. One application, Piggy Bank, extends the Firefox Web browser with various Semantic Web features that transform existing information on the Web (via screen scraping and conversion into RDF) into new combinations and formats not offered by the original Web sites. A faceted user interface allows different filters and display via Google Maps. Various Open Source software is available. It supports a personal tagging 'piggy bank' which can be made public.

5.3.5 SPIRIT

http://www.geo-spirit.org/
SPIRIT (Spatially-Aware Information Retrieval on the Internet) is an EC FP5 research project (partners including University of Cardiff and University of Sheffield) investigating the development of a spatially aware search engine that can automatically recognise geographical terminology. A prototype demo is available. Techniques involve geographic term information extraction and geographic query expansion over spatial relationships and alternate names, with ranked results according to location and topical concepts. It employs a geographical ontology, based on gazetteers and geographical thesauri.

5.3.6 OCLC and OCLC Research

http://www.oclc.org
As owners of the Dewey Decimal Classification, OCLC offer production applications such as WebDewey and the Abridged WebDewey, WorldCat, OCLC Connexion® cataloging interface, etc.

http://www.oclc.org/research/projects/
They also maintain a large research group and OCLC Research are engaged in a variety of vocabulary projects (and related metadata issues such as schema translation), including work on terminology services, persistent identifiers, mapping, automatic classification, schema translation, faceted systems, etc. A variety of datasets and research license software is available.

http://www.oclc.org/research/projects/fast/default.htm
LCSH is a widely used, large, general vocabulary. The FAST project is concerned with developing a simpler syntax while maintaining the richness of the combination of

headings. It has developed automatic methods to de-coordinate LCSH complex headings into individual facets, which are then accessible to processing in their own right. The resulting vocabulary allows more flexible options for indexing and retrieval.

http://www.oclc.org/research/projects/termservices/
They have developed some of the first prototypes in the areas of terminology web services and persistent identifiers. This requires translation of vocabularies from local formats to standard schema, such as MARC 21 or SKOS Core (see Section 6.2). Once in a standard XML-based format, various web services can add value. These include mapping to other vocabularies (see Section 4.4). Mapping services between (pairs of) DDC, LCC, LCSH, MeSH, Eric are offered on the website. Vizine-Goetz *et al.* (2006) report on a prototype that offers an example of an interactive service integrated with desktop applications. Via the Microsoft Office 2003 Research services pane, a user is able copy/paste the results of vocabulary searches into an application in the main window. The intended use is indexing or metadata editing and a choice is available between MARC paste formats and plain text strings for vocabulary terms. SRU/W, REST and SOAP are available as protocols in the underlying web service architecture and full text, SQL or XML storage layers.

A number of vocabularies have been made available on the OCLC Terminology Services Research website, including the following (more are being added):
DCMI Type Vocabulary
DDC Summaries
Library of Congress Subject Headings (LCSH)
Medical Subject Headings (MeSH) 2006
Newspaper Genre List (NGL)

http://www.oclc.org/research/researchworks/ddc/browser.htm
The DeweyBrowser allows search and browsing access via the DDC Summaries (the top 3 levels giving the main structure of the DDC), with multilingual display options. This allows integrated searching and browsing, so that browsing is enabled after an initial search. It can also be used to show the distribution of a collection's records, according to DDC category. An online demonstration is available. The interface is implemented in AJAX (Asynchronous JavaScript and XML).

Recommendation: JISC should negotiate Dewey licenses for JISC services and projects.

5.4 Projects in relation to vocabulary lifecycle framework

In this section, a selection of relevant projects discussed above are roughly placed in relation to the vocabulary lifecycle framework discussed in Section 4.2. Projects are associated by number with points in the framework. Multiple assignments are made and no indication is given of strength of association. Note that this is a very high level categorisation of major emphases in the work and assignment is subjective. Different assignments are possible and this should be viewed as a heuristic for broad overview purposes, not as an evaluative comment on the projects.

Creation and modification of vocabularies
Creating/sustaining vocabularies **5.1.2, 5.1.5, 5.2.4, 5.2.7**

Publication of vocabularies **5.2.7**
Licensing **5.3.6**

Acquisition of vocabularies
selection, storage

Cataloguing (metadata, identification/naming, registration)
Indexing/classification/annotation **5.1.1, 5.1.5, 5.1.7, 5.2.5, 5.3.1, 5.3.3, 5.3.6**
Intellectual, semi-automatic, automatic, disambiguation
5.1.3, 5.2.1, 5.3.2, 5.3.4, 5.3.5, 5.3.6
Integration (syntactic and semantic interoperability issues)
Mapping, merging, linking **5.1.4, 5.2.6, 5.3.3, 5.3.6**

Mediation
User interfaces (TS surfaced in interface) **5.1.5, 5.1.7, 5.2.1, 5.2.5, 5.3.3, 5.3.6**
Faceted/Spatial/Other access afforded **5.1.1, 5.2.2, 5.2.3 / 5.1.1, 5.1.3, 5.3.1, 5.3.5**
Browsing and visualization of vocabularies
Personalization (of interface or vocabulary)

Access, search and discovery
Discovery of services, vocabularies (and concepts), databases/collections, **5.1.4**
Search and retrieval
Querying
TS-aware query **5.1.3, 5.1.4, 5.2.1, 5.2.2, 5.2.6, 5.3.1, 5.3.2, 5.3.5**
Query expansion – synonyms, semantic **5.2.1, 5.2.2, 5.2.6, 5.3.2, 5.3.5**
Cross-searching, cross-browsing across distributed collections
5.1.4, 5.3.2, 5.3.3, 5.3.6

Use (as part of a broader service)
Search+Analysis applications **5.2.6**
Information extraction, mining **5.1.3, 5.3.2, 5.3.4, 5.3.5, 5.3.6**
Application specific, collaborative work

Maintenance of vocabularies
Evolution, versioning

Archiving and preservation of vocabularies

Figure 3. TS Lifecycle framework by project

5.5 Repositories

There is considerable ongoing investment by JISC in repository development, both to enhance open access to content and to improve the management of assets and outputs from the education sector. Whilst this section raises some specific issues that are particularly related to the repository environment as it is now in the UK, the wider aspects of terminology services covered in this review are all relevant to repositories,

Within the UK, repository coverage of research outputs is patchy. Where content is being deposited in institutional repositories, it tends to be from particular departments or subject areas (Heery and Anderson, 2005). Subject coverage is incomplete and unpredictable, and the majority of UK repositories still consist of small collections. In such an environment there needs to be clarity as to the aims of investment in enhancing subject access. There has been some discussion on the American Scientist Open Access Forum mailing list as to the level of end-user subject querying of institutional repositories (Carr, 2006 and following mails), however it is difficult to draw conclusions when repository content and usage is so limited. There is interest in addition of subject terms to content, and guidance to institutional repository administrators is needed to encourage a common approach. The practicality and quality issues that would arise from addition of author (or intermediary) generated subject keywords or classification terms across a distributed base needs to be considered.

Developing scenarios for end-user interaction with repository content would be helpful in order to understand user behaviour (known item searching, author searching etc). A wider perspective on user behaviour would inform how best repository content could be surfaced to meet various search and research requirements. It may be that subject access and terminology services are best considered in relation to aggregated metadata harvested from repositories, and the role of global search engines is fundamental to such an investigation.

Looking ahead to well populated repositories deployed on a wider scale, subject indexing and classification techniques might then underpin various repository services ranging from harvesting by ‘subject sets’ based on OAI-PMH partitions to personalized alerting services using RSS or ATOM. Aggregated metadata harvested from repositories might be automatically classified by subject topic, and other metadata enhancement techniques could be applied. Subject based aggregation of repository content (if IPR issues could be overcome) might provide corpora on which text mining techniques could be applied. The JISC search infrastructure initiative as an aggregator of repository content is already committed to exploring some of these issues, and might work with other specialists to evaluate different approaches.

Repositories should not be seen as isolated ‘silos’ of content, it is important that their content is integrated with other components in the information environment. Whilst subject access to repository content will inevitably have somewhat particular characteristics in the short term, these need to be distinguished from the longer term potential for subject access to repository content and adding value from interfacing with terminology services.

Recommendations:
Pilot different approaches to subject based access to repository content via different types of vocabulary and TS, taking cost benefit issues into account and various levels of aggregation of content:

- **use of subject classification and**
- **use of specialised KOS vocabularies**
- **use of author assigned keywords**
- **full text indexing**

Consider use of mainstream classification (such as DDC) in combination with assigning specialised vocabulary terms (as in use within RDN).

5.6 Augmenting existing programmes and projects

There are many existing JISC programmes and projects, which could be augmented by TS. Some very general scenarios are outlined in Section 2. There is not space in this report to review the broad JISC Service provision, in light of the various research directions described above. There is a need for practical implementations, combined with user evaluations, in order to contribute to a bank of case studies and return on investment data. Various practical recommendations are grouped together in this section.

On the retrieval side, often only uncontrolled keyword search is provided by an information service, without any form of terminology assistance. The range of vocabularies described in Section 3 is available, for consideration of associated cost benefit issues. Basic terminology services, such as synonym expansion, can come at a fairly low overhead and are becoming more common in Web search engines. Sometimes collections are indexed with controlled terminology but this is not systematically taken advantage of on the retrieval side. There are many possibilities for application of query expansion. A simple classification and browsing provision can make a significant difference to user experience. It might sometimes be combined with associated TS, such as lookup over an entry vocabulary, as discussed in Section 2. Integrated search and browsing techniques, as demonstrated by the DeweyBrowser (Section 5.3.6), can be applied to classifications generally. There is scope for work on novel visualizations of vocabulary elements in user interfaces and result displays.

Sometimes no correspondences exist between different parts of the same information service. A basic mapping provision can provide the basis for cross search and cross browsing functionality. While it has still to be investigated, providing more subject awareness to harvesting protocols has potential. Tools are emerging for enhancing the gathering of metadata which could be evaluated in different practical contexts

Recommendations:

- **JISC should support a range of pilot demonstrators with end-users and evaluation**
    - **Investigate different TS approaches to (eg) indexing, mapping, search/browsing, query expansion, disambiguation**

    - **Consider subject access and terminology service adjuncts to appropriate JISC programmes and projects, including TS support for Intute; connection of TS (and subject access) to collection level metadata (e.g. topical composition, correlation); TS support for repositories; project-specific examples.**
- **Harvesting**
    - **Investigate possibilities for extending harvesting tools with more subject metadata**
    - **Investigate relationship of TS and OAI etc**
    - **Evaluate benefits of vocabulary-oriented metadata normalising and enhancement service, e.g. aggregator harvesting relevant metadata, enhancing it and then offering harvesting of the improved metadata**
- **Develop vocabulary visualisation tools supported by TS**
    - **Flexible display and tailoring of segments from vocabularies**
    - **Flexible display and tailoring of results**
    - **Combined search/browsing**

6 Standards

The adoption of common standards for representing and accessing vocabularies has the benefit of enabling interoperability in networked environments and a division of effort. Vocabulary and information resources and searching/indexing/mapping tools may be developed by separate institutions and hosted in separate locations.

Linda Hill and colleagues, discussing the relationship of KOS to Digital Libraries, proposed a service-oriented approach and emphasised the importance of standards:

> Collections, KOSs, and services need to work together in DL architectures. KOSs play a part in collection building, discovery and searching, navigation, evaluation, and visualization. A formal and consistent set of definitions for KOS types, methods for identifying, locating, and referring to individual KOS resources, and protocols for their use will integrate these valuable resources into the overall DL environment. The KOS resources preferred by different communities will be accessible outside of that community for the increasing necessity of cross-domain access to information. The existence of free-standing and accessible KOS resources will counter the tendency to build such systems into particular metadata standards and service protocols.
> Hill *et al.* (2002)

Interoperability requires commonly agreed standards and protocols. Relevant standards, proposed standards and some influential initiatives are briefly reviewed below, with regard to vocabularies and TS generally. It is probably not feasible that any one vocabulary representation schema or access protocol be universally adopted. It is important, however, that vocabulary providers and developers orient to existing standards (and help to evolve them) in the absence of any overriding local imperative. It is also recommended that where possible and appropriate resource providers make available all relevant formats.

Standards exist at different levels and types of interoperability: for design and construction; for representation and interchange; for programmatic access as services.

6.1 Design

Both BSI and NISO have recently published revisions of their standards for thesaurus design, both publications widening their scope and range of vocabularies covered to extend beyond thesauri (BSI, NISO). The BSI Guide is perhaps particularly relevant for JISC UK purposes. Part 2 gives detailed guide to thesauri, facet analysis, display options, including management and planning of thesaurus construction, thesaurus software requirements. Part 3 describes and gives some guidance on vocabularies other than thesauri. Part 4 gives best practice guidance on interoperability and mapping issues, including effect on retrieval. The MARC 21 formats effectively act as standards for various vocabularies, particularly classification schemes. Revised IFLA Guidelines for Multilingual Thesauri have been released for comment (IFLA).

Recommendation: Relevant standards should be included in JISC Standards Catalogue. All new initiatives should take account of relevant design standards

6.2 Representations

In order to support interoperability and service oriented approach, at a syntactical level, representations and interchange formats of vocabularies should be based on XML, unless there is a strong reason otherwise. An XML Schema is probably the simplest option for representing a vocabulary and a variety of other schemes are layered over XML. The MARC 21 Format for Authority Data in XML is now available. The MARC 21 formats are used in DL applications for a variety of vocabularies, as in OCLC's provision. The ZThes 1.0 XML Schema is used by the Zthes profile. The UKgovtalk website provides e-Government Schema Guidelines for XML and links to XML Schemas used in eGov.

XFML is used for some faceted web design applications. VDEX is a proposed standard for eLearning vocabularies. Where compatibility with possible Semantic Web applications is important, RDF/XML should be used. SKOS Core is emerging as an influential RDF representation for vocabularies generally. As a W3C Working Draft, it comes with an extensive guide and documentation and is based on a formal data model. Currently it is under development within the W3C Semantic Web Best Practices and Deployment Working Group working towards Recommendation status. SKOS was originally conceived with thesauri in mind but the scope has been widened to orient to other structured KOS, such as taxonomies and classifications, and less structured vocabularies for social tagging and Web applications.

As mentioned above, in many cases it will be possible and desirable to offer content in a variety of formats and to convert from one format to another, perhaps via XSL and XSLT tools for XML formats. If vocabulary representations are based on an underlying formal model then it is easier to derive transformations to particular syntactical representations. It may sometimes be important to take into account character encodings. For example, Vizine-Goetz *et al.* (2006) discuss issues with automatically converting between MARC-XML and SKOS RDF-XML, where it was found necessary to employ an XSLT 2.0

processor to create an XSL 2.0 transform due to differences in character encodings. Concept or term identifiers may also pose problems since some vocabularies may lack unique IDs or may not have Web actionable URLs (see Section 6.3).

For wider modelling purposes, vocabularies may form an element of a higher level schema and metadata profiles. RDF and OWL should be considered for Semantic Web applications, OWL being appropriate when logical inference is important. The Topic Map XTM standard can be considered for some concept-map and Web-based modelling applications, with a freer structure than formal logic based approaches.

Recommendations:
Strongly recommended to use XML-based representations
Recommended that vocabulary providers consider using SKOS Core if appropriate and contribute to further extensions and customising of SKOS Core

6.3 Identification of concepts, terms and vocabularies

One consequence of operating in a global digital information environment is that the unique identification of resources becomes a major issue. Concepts, metadata terms, vocabularies and the relationships between these various types of entities need to be identified so that they can be automatically referenced and processed. This enables re-use of content and long-term access. In fact, commonly agreed method(s) for handling persistent identifiers are a prerequisite for terminology services, if they are to be interoperable and widely used.

6.3.1 URIs

In the traditional Library world, identifiers such as: the International Standard Book Number (ISBN) and the Book Item Contribution Identifier (BICI) have been used to identify and access resources or their specific parts. Various schemes for identifiers in a networked information environment have been proposed, including the Digital Object Identifier (DOI), Handle, the Uniform Resource Name (URN), Persistent Uniform Resource Locators (PURLS) and the Archival Resource Key (ARK) (cf. the DCC Persistent Identifiers Workshop, 2005). The issue is unlikely to be definitively resolved in the short term - see, for example the report of the ERPANET Seminar on Persistent Identifiers (Simeoni 2004), discussion at the BL/UKOLN March 2006 seminar, 'The Digital Library and its Services' and the NISO Identifiers Roundtable. In the longterm, consideration of identifier options should be placed in the context of an identifier reference model, comprising conceptual, technological, policy, business and social layers (Weibel, cited in Simeoni 2004).

However, there are practical steps which should be followed in current practice. General requirements for identifiers of digital objects (persistence, uniqueness, practicality etc.) should apply to all terminology identifiers, including a capability for immediate dereferencing. *It is recommended that only 'http' URIs currently offer a simple, widely deployed dereferencing mechanism.* It is difficult to see how the non-URI identifier schemes can currently fulfil the necessary conditions for use of terminology identifiers.

On the Web, the Uniform Resource Identifier (URI) provides for the unique identification of resources. It can be used to uniquely identify individual concepts, terms and relationships, so that it is possible to distinguish between entities with the same label.

At the NKOS Special Session at DC 2005, Powell (2005) recommended the use of PURLs as http URIs, allowing dereference to both human-readable and machine-readable term information via an HTTP 303 redirect, and encoding machine-readable information with RDF/RDFS/OWL or SKOS Core. The SKOS Core Guide supports this approach to identifying concepts: "Therefore, the use of any form of HTTP URIs as identifiers for concepts (resources of type skos:Concept) is consistent with the Architecture of the Web, provided that any such resource returns a 303 ('see other') response code in reply to all HTTP GET requests."

Another proposed solution for identifiers is the info-uri scheme, which has been standardised as IETF RFC (Spring 2006). Its purpose is to represent legacy identifiers within a Web context. OCLC have experimented with an info:kos identifier (http://www.oclc.org/research/projects/termservices/resources/info-uri.htm). However, info-uri adds an additional level of indirection in dereferencing. Opinions are divided as to when an agreed mechanism for resolving such identifiers could be established.

6.3.2 Practical experience

OCLC has not taken a long-term decision on identifiers yet, but is experimenting with the use of GUID's (Childress 2005). GUID (Globally Unique Identifier) is a Microsoft implementation of UUID (Universally Unique Identifier) as specified by the Open Software Foundation. They are 16-byte (128-bit) pseudo-random numbers written in hexadecimal. The OCLC Terminology Services project experimentally adds persistent identifiers registered with the info-uri standard (info:kos) to vocabularies, consisting of a scheme and a concept part:

info:kos/scheme/"code"/"expr"/"lang"
info:kos/concept/"code"/"id"

For example, the concept Image in the DCMI Type Vocabulary would be represented info:kos/concept/dct/DCT000004

SKOS Concept identifiers employ URI's, as in the following example:

```
<skos:Concept rdf:about="http://www.example.com/GCL/concepts#529">
    <skos:prefLabel xml:lang="en">Parks and gardens</skos:prefLabel>
    <skos:altLabel xml:lang="en">Allotments</skos:altLabel>
    <skos:altLabel xml:lang="en">Country parks</skos:altLabel>
    <skos:altLabel xml:lang="en">Gardens</skos:altLabel>
    <skos:altLabel xml:lang="en">Grass cutting (garden maintenance)</skos:altLabel>
    <skos:altLabel xml:lang="en">Royal parks</skos:altLabel>
    <skos:broader rdf:resource="http://www.example.com/GCL/concepts#616"/>
    <skos:related rdf:resource="http://www.example.com/GCL/concepts#496"/>
    <skos:related rdf:resource="http://www.example.com/GCL/concepts#1468"/>
    <skos:related rdf:resource="http://www.example.com/GCL/concepts#896"/>
    <skos:inScheme rdf:resource="http://www.example.com/GCL"/>
```

```
    <dct:modified>2002-08-30</dct:modified>
  </skos:Concept>
(see reference for SKOS concept ID)
  <skos:ConceptScheme rdf:about="http://www.example.com/GCL/2.1">
    <dc:title xml:lang="en">Government Category List Version 2.1</dc:title>
    <dc:description xml:lang="en">The GCL (Government Category List) is
      a structured list of categories for use with the Subject.category element
     of the e-GMS.</dc:description>
    <dct:issued>2004-07-01</dct:issued>
(see reference for SKOS ConceptScheme)
```

6.3.3 Further issues

It is important to be clear what the identifier identifies: concept, label (and variants), or record/representation (Childress 2005). In order for an end-user to discover the identifier assigned to a conceptual resource, there is a need for some form of standard registry for the vocabulary (see Section 3.7). Major vocabularies evolve over time and versioning is an ongoing issue. Currently, SKOS does not allow different versions of a KOS to be distinguished (version/edition, language) at the level of concepts. This has been a concern for OCLC and one of the reasons for their experiments with info:kos. The SKOS ConceptScheme information carries version information. The individual concept URI's, however, appear not to do so. Concept scheme versioning is considered an ongoing open issue in the SKOS Core Guide and it is likely there will be further developments.

Recommendations:
A global identifier mechanism for referring to vocabularies and their components underpins interoperable TS.
Recommended to consider building upon existing work with the http URI approach for concept identifiers.
Investigate the addition of identifiers to a widely used freely available vocabulary in a pilot study
Educational work with vocabulary providers on need to supply identifiers and discussions on practical issues should be undertaken

6.4 Protocols, profiles and APIs

Protocols for retrieving vocabulary data are closely linked to representation formats. Generally, a protocol should be defined independently of any particular binding, allowing APIs to be defined for various platforms. It is necessary to distinguish programmatic access to the vocabulary (eg searching or resolving to concepts) from vocabulary support for query (eg as a source of query terms) or browsing.

6.4.1 Protocols to access a vocabulary

There is a role both for general data query protocols and for special vocabulary-based protocols oriented to typical use cases. With regard to the latter, some work has been done on non proprietary terminology protocols for programmatic access to thesauri and related KOS. The HILT project (see Section 5.1.4) has also made use of the commercial WordMap vocabulary API.

The ADL Gazetteer Protocol is emerging as a standard for geographical applications and has been adapted by the Edina projects (see Section 5.1.3). The ADL Thesaurus Protocol offers a lightweight HTTP option for thesaurus access. The protocol's model of a thesaurus closely follows Z39.19 and the definition is specified in an XML schema. A generic, open source Java thesaurus server is supplied and demonstration forms illustrate the five independent services. However, use has mostly been confined within the ADL to date.

Zthes was originally based on Z39.50 but is now available as a profile for SRU and SRW. The Zthes set of specifications includes an Abstract Model for Thesaurus Representation (recently revised to v1.0), an XML Schema and profiles showing "how queries into Zthes-compliant thesauri may be expressed using CQL, and how such thesauri may be accessed using the REST-like SRU protocol and the SOAP-based SRW web-service, or using the ANSI/NISO Z39.50 information retrieval protocol". The Zthes 0.5 Z39.50 protocol has been employed in a few thesaurus Web projects, while the SRW/U Profile is being used by the OCLC Terminology Services project (see Section 5.3.6).

The Simple Knowledge Organisation System (SKOS) API is a recent development, which defines a core set of methods for programmatically accessing and querying vocabularies based on the SKOS-Core RDF schema. While intended as web service calls, the API itself remains independent of implementation details. A web service server implementation has been developed by ILRT and is available for download on an open source basis. One set of SKOS calls returns a concept(s) with its details via an ID, a preferred label, or matching a keyword or regular expression. Another call returns a list of supported semantic relations for the given thesaurus. Another set of calls returns concepts connected by a specified relation or all immediately connected concepts. It is possible also to get a set of concepts connected by a relation up to a given path length.

While few real use interactive applications have been developed with any protocol, the Zthes profile has probably seen the most use and it is integrated with SRW/U, which is widely used. The SKOS API, based on the SKOS RDF representation, has seen less use. It does however include functionality to disclose which relationships are supported rather than hardwiring the thesaurus relationships. It also includes functions that return a composite pattern of concepts and relationships, which is useful for interactive applications. On the other hand, it may be possible to design indexes for the Zthes SRW/U profile, which achieve these composite patterns. The CERES, Zthes 0.5 and ADL protocols were reviewed by Binding and Tudhope (2004), who argued that basing distributed protocol services on atomic vocabulary elements is not necessarily the best approach for applications with interactive user interfaces. Client operations that require multiple client-server calls will carry too much overhead, limiting the responsiveness and interaction styles in the interface. Protocols which offer composite patterns of primitive data elements (via their relationships) may be needed to achieve reasonable response.

While none of the vocabulary protocols reviewed might be considered mature, it is recommended that where possible, projects give thought to adapting an existing protocol.

Experience in real use situations is necessary for further evolution and refinement of the protocols.

6.4.2 Protocols to support query

A vocabulary access protocol can be used in combination with different query protocols. The query protocol may or may not be 'terminology-aware'. A TS might just be a possible source of terms for a free text query. For example, some form of TS might be used in combination with very simple query APIs, such as OpenSearch or SQI. Examples of commercial query APIs include Google, Verity (now Autonomy), FAST. Functionality is evolving rapidly but it is clear that terminology provision is currently included in some of the commercial APIs.

The OCLC Terminology Services project and the DSpace initiative employ SRW/U as the search API. SRW/U is also used by many other DL related projects for basic search operations. While based on the Z39.50 abstract model, it is less complex and is XML based. SRU is a URL REST-based alternative to the SOAP-based SRW. OCLC offers open source client and server software for both versions. CQL is SRW/U's simple but reasonably powerful Boolean query language. CQL itself is not terminology-aware. However, it is possible to define a CQL context set, which has knowledge of vocabulary elements, such as the existing Zthes Thesaurus Context Set for CQL.

Where compatibility with Semantic Web purposes is important then consideration should be given to general semantic query languages, such as SPARQL. SPARQL is a W3C candidate recommendation as a Standard RDF query language. For vocabulary purposes, SPARQL could be considered as both a vocabulary-in-RDF search protocol and as an RDF query protocol. Possibly it might be considered as a lower level protocol upon which some types of application specific protocols could be layered. However, terminology specific protocols will remain important, for applications requiring fast, efficient throughput.

Recommendations:
Need for standard m2m protocols for networked access to vocabularies (and their constituent concepts, relations and terms) with common bindings (APIs) building on web services and other low-level standards
Recommended to consider using SKOS or ZThes API for TS (with a view to contributing to further development). Investigate possibilities of unifying SKOS and ZThes APIs
Investigate possible standard m2m protocols for mapping access to vocabularies, perhaps by expanding SKOS or ZThes APIs
Investigate the combination/integration of TS with existing query APIs (SRU/SRW, CQL) or possibly develop new TS-based query APIs

6.5 Related standards

There are various related standards. See the CETIS website and the JISC Pedagogical Vocabulary Review for discussion of eLearning standards. OCLC research has a list of

standards employed in their developments. Soergel (2001) noted a wide range of standards relevant to KOS.

The Kent State University's Institute for Applied Linguistics maintains a useful set of information on standards related to language technology for translation purposes. Much of this work takes place under the ISO Technical Committee, TC 37 on 'Terminology and other language and content resources'. Standards efforts are involved in various sub-areas: basic principles of terminology management; layout and lexicography; computerized terminology management; natural language processing applications. Work is ongoing on a metadata registry. In many ways, language standards are related to thesaurus standards, but have more emphasis on different word senses, types of term equivalence and uses of terms. For example, instead of the concept versus term distinction, linguistic standards tend to be expressed in terms of three levels: concept - term - lexicalisation as a string.

7 Conclusions

This report has reviewed vocabularies of different types, best practice guidelines, research on terminology services and related projects. It has discussed possibilities for terminology services within the JISC Information Environment and eFramework.

TS can be m2m or interactive, user-facing services and can be applied at all stages of the search process. Services include resolving search terms to controlled vocabulary, disambiguation services, offering browsing access, offering mapping between vocabularies, query expansion, query reformulation, combined search and browsing. These can be applied as immediate elements of the end-user interface or can underpin services behind the scenes, according to context. The appropriate balance between interactive and automatic service components requires careful attention.

Return on investment should be considered in any service provision. There are various types of vocabularies serving different purposes, with different degrees of vocabulary control, richness of semantic relationships, formality, editorial control. There are a range of TS options, both interactive and automatic. There is potential for piloting TS to augment existing JISC programmes and projects.

TS should not be seen as an isolated, free-standing component. TS need to be considered within the wider context of the JISC IE, and need to be integrated with other components of the eFramework. They should be seen as forming a set of services that can be combined with a wide range of other services. There is a need for specifications of TS and their workflow, as part of the JISC IE.

Interoperability requires commonly agreed standards and protocols. Standards exist at different levels and types of interoperability. The prospect is emerging for a broad set of standards across different aspects of terminology services - persistent identifiers, representation of vocabularies, protocols for programmatic access, vocabulary-level metadata in repositories. Such standards are an infrastructure upon which future TS will rest but it is not feasible to wait for international agreements; international consensus will

be influenced by operational experience. Pilot TS projects should orient to existing potential standards (in persistent identifiers, representations, protocols for programmatic access) and help to evaluate and evolve them.

Recent developments have seen a migration of various kinds of services to common desktop applications. The Google toolbar and similar facilities offer rudimentary terminology services. These serve as an inspiration for the more complex types of terminology services that are beginning to be possible.

8 References (by main sections of the review)

3.1 Vocabularies by structure

Adiuri Systems. http://www.adiuri.com/

Aitchison J., Gilchrist A., Bawden D. (2000), Thesaurus construction and use: a practical manual (4th edition), ASLIB, London.

BSI 8723. Structured vocabularies for information retrieval — Guide — Part 1: Definitions, symbols and abbreviations / British Standards Institution. - London : BSI, 2005. - 9p. ; 30cm. - (BS 8723-1:2005) - ISBN 0 580 46798 8.

BSI 8723. Structured vocabularies for information retrieval — Guide — Part 2: Thesauri / British Standards Institution. - London : BSI, 2005. - 59p. ; 30cm. - (BS 8723-2:2005) - ISBN 0 580 46799 6.

BSI 8723. Structured vocabularies for information retrieval — Guide — Part 3: Vocabularies other than thesauri / British Standards Institution. Draft.

BSI 8723. Structured vocabularies for information retrieval — Guide — Part 4: Interoperability between vocabularies / British Standards Institution. Draft.

CIDOC CRM. Conceptual Reference Model. http://cidoc.ics.forth.gr/

Daniels R., Busch J. 2005a. Metadata Best and Worst Practices. Presentation, International Conference on Dublin Core and Metadata Applications, Madrid. Available from http://www.taxonomystrategies.com/html/archive.htm

Daniels R., Busch J. 2005b. Controlled Vocabularies and the Dublin Core. Tutorial, International Conference on Dublin Core and Metadata Applications, Madrid. Available from http://www.taxonomystrategies.com/html/archive.htm

EuroWordNet. http://www.illc.uva.nl/EuroWordNet/

Hearst, Elliott, English, Sinha, Swearingen, and Yee (2002). Finding the Flow in Web Site Search. Communications of the ACM, 45 (9).

Hodge G. 2000. Systems of Knowledge Organization for Digital Libraries: Beyond traditional authority files, Report for The Digital Library Federation Council on Library and Information Resources, 2000. Available online at http://www.clir.org/pubs/abstract/pub91abst.html

NISO - ANSI/NISO Z39.19 - 2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/standard_detail.cfm?std_id=814

Tudhope D., Binding C., Blocks D., Cunliffe D. 2006. Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509-533.

WordNet. http://wordnet.princeton.edu/w3wn.html

Yee, K-P., Swearingen, K., Li, K., Hearst, M. (2003), "Faceted Metadata for Image Search and Browsing", Proc. ACM Conference on Human Factors in Computing Systems, pp. 401-408.

3.2 Vocabularies by purpose

Gruber T. What is an ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
Lancaster F. 2003. Indexing and Abstracting in Theory and Practice. Facet (3rd edition).
Middleton's Controlled Vocabulary List.
Smith B. 2003. Ontology. In: (L. Floridi (ed.), Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell, 2003, 155–166.

3.2.4 eLearning purposes - Indicative eLearning Vocabulary-related References

(See also JISC Pedagogical Vocabularies Project and its List of Vocabularies)

Guides and Studies

Barker P. 2005. What is IEEE Learning Object Metadata / IMS Learning Resource Metadata? CETIS Guide. http://metadata.cetis.ac.uk/guides/#What_is...
Currier, S., Barton, J., O'Beirne, R., Ryan, B. 2004. Quality assurance for digital learning object repositories: issues for the metadata creation process. ALT-J : Research in Learning Technology, 12(1), 5-20.
CETIS Briefings http://www.cetis.ac.uk/static/briefings.html
CETIS Guides http://metadata.cetis.ac.uk/guides/
Fegen N. 2006. CETIS Briefing paper on VDEX. http://metadata.cetis.ac.uk/guides/WhatIsVDEX.pdf
JISC Pedagogical Vocabularies Project, with 3 Reports. http://www.jisc.ac.uk/elp_vocabularies.html
Powell A., Barker P. 2004. RDN/LTSN Partnerships: Learning resource discovery based on the LOM and the OAI-PMH. Ariadne 39, April, 2004. http://www.ariadne.ac.uk/issue39/powell/
Smith T., Zeng M. 2004. Building Semantic Tools for Concept-based Learning Spaces: Knowledge Bases of Strongly-Structured Models for Scientific Concepts in Advanced Digital Libraries. Journal of Digital Information, 4(4), Article No. 263, 2004-01-28. http://jodi.tamu.edu/Articles/v04/i04/Smith/

Guidelines

CanCore Guidelines: Access For All Digital Resource Description data elements. http://www.cancore.ca/en/guidelines.html
IMS Meta-data Best Practice Guide for IEEE 1484.12.1-2002 Standard for Learning Object Metadata - Version 1.3 Public Draft
http://www.imsglobal.org/metadata/mdv1p3pd/imsmd_bestv1p3pd.html
RLLOMAP: A cataloguer's handbook. 2nd edition: May 2005. http://www.heacademy.ac.uk/Cataloguers_handbook2.doc

Some projects and organizations

ARIADNE. http://www.ariadne-eu.org

Becta Vocabulary Studio and Becta Vocabulary Bank.
http://www.becta.org.uk/vocab/index.cfm http://www.becta.org.uk/vocab/browse.cfm
CETIS - http://www.cetis.ac.uk/
EdNA. http://www.edna.edu.au
JORUM. http://www.jorum.ac.uk/
MERLOT. http://taste.merlot.org
RDN. http://www.rdn.ac.uk
SchemaLogic. www.schemalogic.com
Vocabulary Management Group. http://www.vocman.com/

Models and Application Profiles
DC-Education Application Profile and DCMI Education Working Group.
http://dublincore.org/groups/education/
IMS Vocabulary Definition Exchange (VDEX) specification.
http://www.imsglobal.org/vdex/index.html
RDN/LTSN LOM Core Application Profile. The Higher Education Academy.
http://www.rdn.ac.uk/publications/rdn-ltsn/ap/
UK Learning Object Metadata Core (UK LOM Core).
http://www.cetis.ac.uk/profiles/uklomcore

Tagging tools
Becta Tagging Tool. http://www.becta.org.uk/vocab/index.cfm
Curriculum Online. http://www.curriculumonline.gov.uk/SupplierCentre/taggingtool.htm
RELOAD. http://www.reload.ac.uk

Indicative examples of eLearning Vocabularies by Category
Educational Objective
LTSN pedagogic terms vocabulary. http://www.rdn.ac.uk/publications/rdn-ltsn/pedagogic-terms/
Bloom B. 1956. (ed.) Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: Cognitive Domain. New York: Longman.

Educational level
UK Educational Levels vocabulary (UKEL).
http://www.ukoln.ac.uk/metadata/education/ukel/

Discipline (general vocabulary)
DDC Dewey Decimal Classification
LCSH Library of Congress Subject Headings.
UDC Universal Decimal Classification

Idea (subject specific vocabulary)
Art and Architecture Thesaurus.
http://www.getty.edu/research/conducting_research/vocabularies/aat/index.html
British Education Thesaurus. http://brs.leeds.ac.uk/~beiwww/beid.html

CAB Thesaurus. http://www.cabi-publishing.org/DatabaseSearchTools.asp?SubjectArea=&PID=277
HASSET (Humanities And Social Sciences Electronic Thesaurus). http://www.data-archive.ac.uk/search/hassetSearch.asp
Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html

Idea (curriculum related vocabulary)
Joint Academic Coding System http://www.hesa.ac.uk/jacs/jacs.htm
Learndirect Classification System http://www.learndirect-advice.co.uk/provider/standardsandclassifications/classpage/

Resource type
RDN/LTSN resource type vocabulary. http://www.rdn.ac.uk/publications/rdn-ltsn-ap/types/

3.3 Named entity authority and disambiguation services

Armadillo Project, University of Sheffield. http://www.hrionline.ac.uk/armadillo/sources.html
Ask/Bloglines Blog and Feeds Search http://www.bloglines.com/
Crane G., Jones A. 2006. The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. Proc.. JCDL 2006, Chapel Hill, ACM Press.
Crane G. 2004. Georeferencing in Historical Collections. D-Lib Magazine, 10(5), May 2004. http://www.dlib.org.ar/dlib/may04/crane/05crane.html
D2K http://alg.ncsa.uiuc.edu/do/tools/d2k
Fayyad U., Grinstein G., Wierse A. 2001. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publ.
GATE project. University of Sheffield. http://gate.ac.uk/
GeoCrossWalk. URL: http://www.geoxwalk.ac.uk/
GBHGIS http://www.geog.port.ac.uk/gbhgis/
GRADE project http://edina.ac.uk/projects/grade
Guy M., Powell A., Day M. 2004. Improving the Quality of Metadata in Eprint Archives. Ariadne 38, Jan. 2004 http://www.ariadne.ac.uk/issue38/guy/
Han H., Zha, H., Giles C. 2005. Name disambiguation in author citations using K-way spectral clustering method. Proc. JCDL 2005, 334-343. ACM Press.
HEIRPORT Project. http://ads.ahds.ac.uk/heirport/
IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). http://www.ifla.org/VII/d4/wg-franar.htm
Hill L. 2004. (Guest ed.) Georeferencing in Digital Libraries. Special Issue, DLIB Magazine, 10(5), May 2004. http://www.dlib.org.ar/dlib/may04/05contents.html
Hillmann D., Dushay N., Phipps J. 2004. Improving Metadata Quality: Augmentation and Recombination. Proc. DC2004 Conference, Shanghai. http://metamanagement.comm.nsdl.org/Metadata_Augmentation-DC2004.html
LC Authorities http://authorities.loc.gov/
LC Name Authority File Web service http://alcme.oclc.org/eprintsUK/
LEAF project http://www.crxnet.com/leaf/

Lynch C. 2006. Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures. http://www.cni.org/staff/clifford_publications.html
NaCTeM. National Centre for Text Mining. http://www.nactem.ac.uk/
OCLC Metadata Switch project http://www.oclc.org/research/projects/mswitch/default.htm
NORA project http://www.noraproject.org
NSDL. http://nsdl.org/
Perseus project http://www.perseus.tufts.edu/
Petras V., Larson R., Buckland M. 2006. Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context. Proc.. JCDL 2006, Chapel Hill.
Shneiderman B. 2002. Inventing discovery tools: Combining information visualization with data mining. In: Information Visualization 1, 1, 5-12.
Smith D. 2002. Detecting events with date and place information in unstructured text. In: Proc. JCDL 2002, 191-196. ACM Press
STLQ http://stlq.info/2006/05/scopus_author_identifier_new_f.html
VIAF http://www.oclc.org/research/projects/viaf/default.htm
VISION http://www.visionofbritain.org.uk/expertsearch.jsp
Voss, J. 2005. Metadata with Personendaten and beyond. Proc. 1st Wikimania Conference, http://meta.wikimedia.org/wiki/Transwiki:Wikimania05/Paper-JV2
Wikipedia disambiguation http://en.wikipedia.org/wiki/Wikipedia:Disambiguation
Witten I., Frank E. 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Academic Press.

3.4 Social tagging and folksonomies

CiteUlike http://www.citeulike.org/
Connotea http://www.connotea.org/
Davis I. 2005. Why tagging is expensive. http://silkworm.talis.com/blog/archives/2005/09/why_tagging_is.html
Delicious http://del.icio.us/
Digg http://digg.com/
fac.etio.us, Siderean's faceted search of delicious tags http://www.siderean.com/facetious/facetious.jsp (no longer available, 28 June 2006)
Flickr http://www.flickr.com/
Folksonomy. http://en.wikipedia.org/wiki/Folksonomy
Golder S., Huberman B. 2006. Usage patterns of collaborative tagging systems. Journal of Information Science 32:2, 198-208
Hammond T., Hannay T., Lund B., Scott J. Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11(4), 2005. http://www.dlib.org/dlib/april05/hammond/04hammond.html
Hannay T. Introduction. August 19, 2004. http://tagsonomy.com/index.php/introduction-timo-hannay/
Last.fm http://www.last.fm/
Lund B., Hammond T., Flack M., Hannay T. 2005. Social Bookmarking Tools (II): A Case Study - Connotea. D-Lib Magazine, 11(4), 2005. http://www.dlib.org/dlib/april05/lund/04lund.html

Tagging. http://en.wikipedia.org/wiki/Tagging
Technorati http://www.technorati.com/
Trant J. 2006. Exploring the potential for social tagging and folksonomy in art museums: proof of concept. New Review of Hypermedia and Multimedia, 12(1), 83-105.

3.5 Best practice guidelines for constructing and using vocabularies

Aitchison J., Gilchrist A., Bawden D. 2000. Thesaurus construction and use: a practical manual (4th edition), ASLIB, London.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 1: Definitions, symbols and abbreviations / British Standards Institution. - London : BSI, 2005. - 9p. ; 30cm. - (BS 8723-1:2005) - ISBN 0 580 46798 8.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 2: Thesauri / British Standards Institution. - London : BSI, 2005. - 59p. ; 30cm. - (BS 8723-2:2005) - ISBN 0 580 46799 6.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 3: Vocabularies other than thesauri / British Standards Institution. Draft.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 4: Interoperability between vocabularies / British Standards Institution. Draft.
Daniels R., Busch J. 2005a. Metadata Best and Worst Practices. Presentation, International Conference on Dublin Core and Metadata Applications, Madrid. Available from http://www.taxonomystrategies.com/html/archive.htm
Daniels R., Busch J. 2005b. Controlled Vocabularies and the Dublin Core. Tutorial, International Conference on Dublin Core and Metadata Applications, Madrid. Available from http://www.taxonomystrategies.com/html/archive.htm
Dextre Clarke. Taxonomies and thesauri: a list of references and resources for public sector applications. www.govtalk.gov.uk/documents/Bibliography2005-05-11.rtf
e-Government Metadata Standard (Version 3.1). Available from http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=1017
English Heritage NMR thesauri. http://thesaurus.english-heritage.org.uk/frequentuser.htm
GovTalk Archive. Design/selection criteria for software used to handle controlled vocabularies. http://www.govtalk.gov.uk/archive/archive.asp?librarydocs=3
Lancaster F. 2003. Indexing and Abstracting in Theory and Practice. Facet (3rd edition).
Middleton's Controlled Vocabulary List. http://sky.fit.qut.edu.au/%7Emiddletm/cont_voc.html
NISO - ANSI/NISO Z39.19 - 2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/standard_detail.cfm?std_id=814
University of British Columbia. Indexing Resources on the WWW. Database Indexing, Controlled Vocabularies & Thesauri. http://www.slais.ubc.ca/resources/indexing/database1.htm
University of Toronto Library's Subject Analysis Systems (SAS) Collection. This contains information on how to search the SAS via the Library Catalogue http://www.fis.utoronto.ca/content/view/386/134/
WillPower Information. Publications on thesaurus construction and use. Links to lists of thesauri. http://www.willpower.demon.co.uk/thesbibl.htm
WillPower Thesaurus Software. http://www.willpower.demon.co.uk/thessoft.htm

3.6 Network access to Vocabularies

Dextre Clarke. Taxonomies and thesauri: a list of references and resources for public sector applications. www.govtalk.gov.uk/documents/Bibliography2005-05-11.rtf
English Heritage NMR thesauri. http://thesaurus.english-heritage.org.uk/frequentuser.htm
HILT A-Z of Thesauri. http://hilt.cdlr.strath.ac.uk/hilt2web/Sources/thesauri.html
Koch's Controlled vocabularies, thesauri and classification systems available in the WWW. DC Subject. http://www.lub.lu.se/metadata/subject-help.html
MDA Terminology Bank. http://www.mda.org.uk/spectrum-terminology/termbank.htm
Species 2000. http://www.sp2000.org/
SWAD-Europe Project list of thesauri (no longer maintained). http://www.w3.org/2001/sw/Europe/reports/thes/thes_links.html
TASI links to metadata vocabularies. http://www.tasi.ac.uk/resources/vocabs.html
Taxonomy Warehouse. http://www.taxonomywarehouse.com/
Text Mining Centre. http://www.nactem.ac.uk/
University of British Columbia. Indexing Resources on the WWW. Database Indexing, Controlled Vocabularies & Thesauri. http://www.slais.ubc.ca/resources/indexing/database1.htm
University of Toronto Library's Subject Analysis Systems (SAS) Collection. This contains information on how to search the SAS via the Library Catalogue http://www.fis.utoronto.ca/content/view/386/134/
WillPower Information. Publications on thesaurus construction and use. Links to lists of thesauri. http://www.willpower.demon.co.uk/thesbibl.htm

3.7 Terminology Services Registries

GRIMOIRES Grid Registry with Metadata Oriented Interface: Robustness, Efficiency, Security; University of Southampton Electronics and Computer Science. http://www.ecs.soton.ac.uk/research/projects/grimoires
ISO/IEC 11179, Information Technology -- Metadata Registries (MDR) http://metadata-standards.org/11179/
XMDR Extended Metadata Registry Project http://www.xmdr.org/

4.1 Studies and models of information seeking behaviour

Bates M. 1979. Information search tactics. Journal of the American Society for Information Science, 30(4), 205-214.
Bates M. 1989. The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(5), 407-424.
Bates M. 1990. Where should the person stop and the information search interface start? Information Processing & Management. 26(5), 575-591.
Beaulieu M. 1997. Experiments on interfaces to support query expansion. Journal of Documentation, 53(1), 8-19.
Blocks D., Cunliffe D. Tudhope D. 2006. A reference model for user-system interaction in thesaurus-based searching. Journal of the American Society for Information Science and Technology, 57, (in press).

Brajnik G., Mizzaro S., Tasso C. 1996. Evaluating user interfaces to information retrieval systems: A case study on user support. Proc.19th ACM SIGIR conference, 128-136.
Choo C., Detlor B., Turnbull D. 2000. Information seeking on the Web: An Integrated Model of Browsing and Searching. First Monday, 5(2). http://www.firstmonday.dk/issues/issue5_2/choo/index.html
Ellis D. 1989. A behavioural approach to information retrieval systems design. Journal of Documentation, 45(3), 171-212.
Fidel R. 1985. Moves in online searching. Online Review, 9(1), 61-74.
Fidel R. 1991. Searchers' selection of search keys (I-III), Journal of the American Society for Information Science, 42(7), 490-527.
Fidel R., Efthimiadis E.1995. Terminological knowledge structure for intermediary expert-systems. Information Processing & Management, 31(1), 15-27.
Greenberg J. 2001. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. Journal of the American Society for Information Science, 52(6), 487-498.
Jones S., Gatford M., Robertson, S., Hancock-Beaulieu, M., Secker, J., Walker, S. 1995, Interactive Thesaurus Navigation: Intelligence Rules OK? Journal of the American Society for Information Science, 46(1), 52-59.
Kuhlthau C. 1991. Inside the search process - information seeking from the users perspective. Journal of the American Society for Information Science, 42(5), 361-371.
Marchionini G. 1995. Information seeking in electronic environments. Cambridge: Cambridge University Press.
Shiri A., Revie C. 2006. Query expansion behaviour within a thesaurus-enhanced search environment: a user-centered evaluation. Journal of the American Society for Information Science, 57(4), 462–478.
Soergel D. 1994. Indexing and retrieval performance: The logical evidence, Journal of the American Society for Information Science, 45(8), 589-599.
Spink A., Wilson T., Ford N., Foster A., Ellis D. 2002. Information-seeking and mediated searching. Journal of the American Society for Information Science and Technology, 53(9), 695-703.
Wilson, T. D. 1999. Models in information behaviour research. Journal of Documentation, 55(3), 249-270.
Vakkari P., Jones S., MacFarlane A., Sormunen E. 2004. Query exhaustivity, relevance feedback and search success in automatic and interactive query expansion, Journal of Documentation, 60 (2), 109-127.

4.2 Information lifecycle with regard to TS

Patel M., Koch T., Doerr M., Tsinaraki C. 2005. Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1.
Lyon L. 2003. eBank UK: Building the links between research data, scholarly communication and learning. Ariadne, July 2003, Issue 36. http://www.ariadne.ac.uk/issue36/lyon/

4.3 Types of Terminology Web Services

Binding C., Tudhope D. 2004. KOS at your Service: Programmatic Access to Knowledge Organisation Systems", Journal of Digital Information, Volume 4, Issue 4. http://jodi.tamu.edu/Articles/v04/i04/Binding/

Blocks D., Cunliffe D. Tudhope D. 2006. A reference model for user-system interaction in thesaurus-based searching. Journal of the American Society for Information Science and Technology, 57, (in press).

CalEDLN Thesaurus API http://northbaycommons.net/the-developers-group/developers-wiki/ceres-api/ Thesaurus Browser: http://ceres.ca.gov/search/

CSA/NBII Biocomplexity thesaurus web services. http://nbii-thesaurus.ornl.gov/thesaurus/

DARE http://www.darenet.nl/en/page/language.view/home

DLF Abstract Services Taskforce http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm

e-Framework for Education and Research. http://www.e-framework.org/

ELF E-Learning Framework http://elframework.org/

GEMET web services. http://www.eionet.europa.eu/gemet/webservices?langcode=en

Melvil http://www.ddc-deutsch.de/literature/2005_3_Melvil.pdf

MeSHine http://www.meshine.info

LC Name Authority File Web Service http://alcme.oclc.org/eprintsUK/

OCLC Terminology Services Project and Pilot http://www.oclc.org/research/projects/termservices/ http://www.oclc.org/research/projects/termservices/resources/tspilot-services.htm

Powell A. 2005a, Feb. JISC IE Discovery to Delivery (D2D) Reference Model. Draft for discussion. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/dlf/

Powell A. 2005b, Nov. A 'service oriented' view of the JISC Information Environment. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf

SKOS API. http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html

Sumner, T. 2005. The NSDL Strand Map Service: A Networked Knowledge Organization and Visualization System for K-12 Education. Presentation NKOS workshop at JCDL 2005. http://nkos.slis.kent.edu/2005workshop/sumner.ppt

Tudhope D., Binding C. 2006. Toward Terminology Services: Experiences with a Pilot Web Service Thesaurus Browser, ASIS&T Bulletin, June/July, 2006. Available online at http://www.asist.org/Bulletin/Jun-06/tudhope_binding.html

Vizine-Goetz D, Hickey C, Houghton A, Thompson R. 2003. Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4), Article No. 272, 2004-03-11. http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/

Vizine-Goetz D., Houghton A., Childress E. 2006. "Web Services for Controlled Vocabularies", ASIS&T Bulletin, June/July 2006, Available online at http://www.asist.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html (accessed 10 June 2006).

W3C Semantic Web Services Interest Group http://www.w3.org/2002/ws/swsig/

W3C Web Services http://www.w3.org/2002/ws/

Zisman, A., Chelsom, J., Dinsey, N., Katz, S. and Servan, F. 2002. Using Web Services to Interoperate Data at the FAO. Proc. International Conference on Dublin Core and Metadata for e-Communities (Firenze UP), 147-156

4.4 Mapping

BSI 8723. BSI Standard 8723 on Structured Vocabularies for Information Retrieval – Guide. Part 4: Interoperability between Vocabularies. British Standards Institution. - London : BSI, 2006.

Doerr M. 2001. Semantic Problems of Thesaurus Mapping, Journal of Digital Information, 1(8), Article No. 52, 2001-03-26. http://jodi.tamu.edu/Articles/v01/i08/Doerr/

Doerr M., Hunter J., Lagoze C. (2003), "Towards a Core Ontology for Information Integration", Journal of Digital Information, 4 (1), http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/

Koch, T., Neuroth, H., Day, M. 2003. Renardus: Cross-browsing European subject gateways via a common classification system (DDC). Proc. IFLA Satellite Meeting 2001, (ed. I. McIlwaine), IFLA UBICIM Publ., New Series 25 (München: K G Saur), 25-33 - preprint http://www.lub.lu.se/~traugott/drafts/preifla-final.html

Liang A., Sini M. 2006. Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review of Hypermedia and Multimedia, 12(1), 51-62.

Navarretta C., Pedersen B., Hansen D. 2006. Language technology in knowledge organization systems. New Review of Hypermedia and Multimedia, 12(1), 29-49.

OCLC Terminology Services Research Project. http://www.oclc.org/research/projects/termservices/default.htm

Patel M., Koch T., Doerr M., Tsinaraki C. 2005. Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1.

SKOS Mapping Schema. http://www.w3.org/2004/02/skos/mapping/

Vizine-Goetz D, Hickey C, Houghton A, Thompson R. 2003. Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4), Article No. 272, 2004-03-11. http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/

Zeng M, Chan L. 2004. Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377 – 395.

4.5 Automatic classification and indexing

ALVIS Project. http://www.alvis.info/alvis/

Combine tool. http://combine.it.lth.se/

Collexis. http://www.collexis.com and for licenses for developing countries see IntellectuALL http://www.intellectuall.org/home/modules/tinycontent/?id=1

DESIRE. DESIRE Project. http://www.desire.org

Golub K. 2006. Automated subject classification of textual Web documents. Journal of Documentation, 62 (3), pp. 350-371.

Golub. K. 2006. Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations. New Review of Hypermedia and Multimedia, 12(1), 11-27

Godby J., Stuler J. 2001. The Library of Congress Classification as a knowledge base for automatic subject categorization. Presentation IFLA Preconference, Subject Retrieval in a Networked Environment, http://staff.oclc.org/~godby/auto_class/godby-ifla.html

Hagedorn K. 2001. Extracting Value from Automated Classification Tools: The Role of Manual Involvement and Controlled Vocabularies. ACIA White Paper. http://argus-acia.com/white_papers/classification.html
iVia project and tools. http://ivia.ucr.edu/
JCDL 2006 Metadata tools for digital resource repositories workshop
Exhibitors. http://www.ils.unc.edu/mrc/jcdl2006/MetadataWorkshopExhibitors.pdf
Koch T., and Ardö A., 2000. Automatic classification of full-text HTML-documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2. Available online at: http://www.lub.lu.se/desire/DESIRE36a-WP2.html
Koch T., Neuroth H., Day M. 2003. Renardus: Cross-browsing European subject gateways via a common classification system (DDC). Proc. IFLA Satellite Meeting 2001, (ed. I. McIlwaine), IFLA UBICIM Publications, New Series 25 (München: K G Saur), 25-33 - preprint http://www.lub.lu.se/~traugott/drafts/preifla-final.html
KnowLib Demonstrators and tools
http://www.it.lth.se/knowlib/auto.htm
http://www.it.lth.se/knowlib/demos.htm
Lancaster F. 2003. Indexing and Abstracting in Theory and Practice. Facet (3rd edition).
Larson, R.R. 1992. Experiments in automatic Library of Congress Classification. Journal of the American Society for Information Science, 43(2), 130-148
Middleton's Controlled Vocabulary List.
http://sky.fit.qut.edu.au/%7Emiddletm/cont_voc.html
Medelyan, O. and Witten, I. 2006. Thesaurus Based Automatic Keyphrase Indexing. Proc. JCDL 2006, 296-297.
OCLC Automatic Classification project.
http://www.oclc.org/research/projects/auto_class/default.htm
OCLC Scorpion. http://www.oclc.org/research/software/scorpion/
OCLC Connexion. http://www.oclc.org/connexion/
Paynter G. 2005. Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources. Proc. JCDL 2005, ACM Press, 291-300.
Russell R., Day. M. 2001. Automatic indexing and classification tools. Review for HILT Project. http://www.ukoln.ac.uk/metadata/hilt/interfaces/

4.6 Text mining and information extraction

Ananiadou S., Chruszcz J., Keane J., McNaught J., Watry P. 2005. The National Centre for Text Mining: Aims and Objectives, Ariadne, 42. Jan. 2005.
NaCTeM. National Centre for Text Mining. http://www.nactem.ac.uk/
Lynch, Clifford. Open computation: beyond human-reader-centric views of scholarly literatures, *in* Jacobs, N., (Ed.) Open access: key strategic, technical and economic aspects, Chandos Publishing, 2006.
http://www.cni.org/staff/cliffpubs/OpenComputation.htm

4.7 General sources for work in TS

L. Hill and T. Koch, Networked Knowledge Organization Systems: introduction to a special issue, Journal of Digital Information, 1(8), Article No. 53, 2001-04-03, 2001. Available online at http://jodi.tamu.edu/Articles/v01/i08/editorial/

NKOS Network, Networked Knowledge Organization Systems/Services. http://nkos.slis.kent.edu/ (accessed 10 June 2006).
D. Tudhope, T. Koch, New Applications of Knowledge Organization Systems: introduction to a special issue, Journal of Digital Information, 4(4), Article No. 286, 2004-02-13, 2004. Available online at http://jodi.tamu.edu/Articles/v04/i04/editorial/
D. Tudhope and Nielsen M. 2006. Introduction to Special Issue on Knowledge Organization Systems and Services. New Review of Hypermedia and Multimedia, 12(1), 3-9.

5.5 Repositories

Carr, Leslie. E-Mail: Use of Navigational Tools in a Repository http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/5170.html
Heery, R. and Anderson, S. Digital repositories review. Report to accompany JISC Digital Repositories Programme call, February 2005. http://www.jisc.ac.uk/index.cfm?name=programme_digital_repositories

6 Standards

L. Hill, O. Buchel, G. Janée and M. Zeng, "Integration of Knowledge Organization Systems into Digital Library Architectures", Position Paper 13th ASIS&T SIG/CR Workshop, Reconceptualizing Classification Research, 2002. Available online at http://www.alexandria.ucsb.edu/~gjanee/archive/2002/kos-dl-paper.pdf

6.1 Design

ANSI/NISO Z39.19 - 2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/standard_detail.cfm?std_id=814
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 1: Definitions, symbols and abbreviations / British Standards Institution. - London : BSI, 2005. - 9p. ; 30cm. - (BS 8723-1:2005) - ISBN 0 580 46798 8.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 2: Thesauri / British Standards Institution. - London : BSI, 2005. - 59p. ; 30cm. - (BS 8723-2:2005) - ISBN 0 580 46799 6.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 3: Vocabularies other than thesauri / British Standards Institution. Draft.
BSI 8723. Structured vocabularies for information retrieval — Guide — Part 4: Interoperability between vocabularies / British Standards Institution. Draft.
IFLA. Revised IFLA Guidelines for Multilingual Thesauri - released for comment. http://www.ifla.org/VII/s29/wgmt-invitation.htm

6.2 Representations

ADL XML Thesaurus Schema http://www.alexandria.ucsb.edu/thesaurus/protocol/thesaurus-protocol.xsd
MARC 21 formats. http://www.loc.gov/marc/marcdocz.html
MARC 21 XML Schema http://www.loc.gov/standards/marcxml/
XFML eXchangeable Faceted Metadata Language. http://xfml.org/
SKOS Core

http://www.w3.org/2004/02/skos/
http://www.w3.org/2004/02/skos/core/
http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
Topic maps and XML Topic Maps
http://www.techquila.com/topicmaps.html
http://www.topicmaps.org/
http://www.ontopia.net/topicmaps/materials/tao.html
UKgovtalk e-Government Schema Guidelines for XML
http://www.govtalk.gov.uk/schemasstandards/developerguide_document.asp?docnum=94
VDEX - IMS Vocabulary Definition Exchange (VDEX) specification.
http://www.imsglobal.org/vdex/index.html
Vizine-Goetz D., Houghton A., Childress E. 2006. Web Services for Controlled Vocabularies, ASIS&T Bulletin, June/July 2006, Available online at http://www.asist.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html
Zthes. http://zthes.z3950.org/

6.3 Identification of concepts, terms and vocabularies

Childress, E., Houghton, A. and Vizine-Goetz, D. 2005. OCLC and vocabulary identifiers. Presentation at the NKOS Special Session, DC 2005, Madrid. http://www.ukoln.ac.uk/terminology/events/NKOSatDC2005/OCLC and vocabulary identifiers.ppt
DCC Persistent Identifiers Workshop, Univ. of Glasgow, June 2005.
http://www.dcc.ac.uk/training/pi-2005/
Info-uri IETF RFC http://www.ietf.org/rfc/rfc4452.txt
Powell A. 2005. (Persistent) Identifiers for Concepts / Terms / Relationships. Presentation at the NKOS Special Session, DC 2005, Madrid. http://www.ukoln.ac.uk/terminology/events/NKOSatDC2005/Powell-persistent-identifiers.ppt
Simeoni, F. 2004. A report on the ERPANET Seminar on Persistent Identifiers, 17-18 June 2004, Cork, Ireland. http://hairst.cdlr.strath.ac.uk/documents/Erpanet%20Training%20Seminar%20on%20Persistent%20Identifiers.pdf
SKOS concept ID example
http://www.w3.org/2004/02/skos/core/examples/Concept.rdf.xml
SKOS ConceptScheme example
http://www.w3.org/2004/02/skos/core/examples/ConceptScheme.rdf.xml
W3C: Naming and Addressing: URIs, URLs, ... http://www.w3.org/Addressing/

6.4 Protocols, Profiles and APIs

Binding C., Tudhope D. 2004. KOS at your Service: Programmatic Access to Knowledge Organisation Systems, Journal of Digital Information, 4(4), Article No. 265, 2004-02-05 http://jodi.tamu.edu/Articles/v04/i04/Binding/
FAST API. http://www.fastsearch.com/press.aspx?m=63&amid=565
Google API http://code.google.com/apis.html

OpenSearch API http://opensearch.a9.com/ (simple API)
http://www.unto.net/unto/work/on-open-search-apis/
SKOS API. http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
SQI API – Simple Query Interface (Simple eLearning API)
http://www.cs.kuleuven.ac.be/~hmdb/ProlearnIClass/papers/Ternier.htm
http://ariadne.cs.kuleuven.ac.be/vqwiki-2.5.5/jsp/Wiki?LorInteroperability
Verity http://www.autonomy.com/content/News/Releases/2003/V0908a.html
WordMap http://www.wordmap.com/index.html
Zthes. http://zthes.z3950.org/

SRU/W
http://www.oclc.org/research/projects/webservices/default.htm
http://www.loc.gov/standards/sru/ SRU/W
http://www.loc.gov/standards/sru/cql/index.html CQL

W3C
Semantic Web http://www.w3.org/2001/sw/
Semantic Web Services Interest Group http://www.w3.org/2002/ws/swsig/
SPARQL http://www.w3.org/TR/rdf-sparql-query/
RDF http://www.w3.org/RDF/
OWL http://www.w3.org/TR/owl-guide/
Web Services http://www.w3.org/2002/ws/

6.5 Related Standards

ISO Technical Committee TC 37. Terminology and other language and content resources
http://www.iso.org/iso/en/CatalogueListPage.CatalogueList?COMMID=1459&scopelist=PROGRAMME
Kent State University's Institute for Applied Linguistic.
http://appling.kent.edu/ResourcePages/LTStandards/Chart/LanguageTechnologyStandards.htm
Soergel D. 2001. The representation of Knowledge Organization Structure (KOS) data: a multiplicity of standards. JCDL 2001 NKOS Workshop, Roanoke.
http://www.dsoergel.com/cv/B75.pdf